人工智能安全风险测评

（2025 年）

——白皮书——

中国信息安全测评中心
China Information Technology Security Evaluation Center

2025 年 12 月

前 言

当前，全球人工智能迅速发展，已经成为引领新一轮科技革命和产业变革的战略力量，以前所未有的广度与深度重塑社会生产生活方式，为高质量发展注入新的动能。与此同时，人工智能广泛应用带来前所未有的安全挑战，潜藏多重复杂的威胁风险。世界各国和地区组织高度重视人工智能安全，围绕加强人工智能安全与监管，纷纷采取多样性、差异化的管理措施，竞争人工智能发展和治理主动权。

我国2023年提出《全球人工智能治理倡议》，明确“坚持发展和安全并重的原则，构建开放、公正、有效的治理机制”，“推动建立风险等级测试评估体系”“打造可审核、可监督、可追溯、可信赖的人工智能技术”；2025年发布《关于深入实施“人工智能+”行动的意见》，要求“提升安全能力水平”“推动人工智能应用合规、透明、可信赖”“建立健全人工智能技术监测、风险预警、应急响应体系”“加快形成动态敏捷、多元协同的人工智能治理格局。”

《人工智能安全风险测评（2025年）》白皮书结合国内人工智能领域战略规划、政策法规、标准体系等基础，探讨人工智能安全风险测试评估的实施路线，推动人工智能产业形成“测

评—反馈—迭代”安全闭环，促进人工智能技术实现“安全可控”与“创新发展”动态平衡，服务人工智能发展与监管。

在内容安排方面，白皮书围绕“为何测（Why）”的核心目标，通过总结分析梳理，推动“安全性、可靠性、可控性、公平性”及“可审核、可监督、可追溯、可信赖”等治理要求转化为可落地的人工智能安全风险工作流程；解析人工智能安全风险分析框架，绘制“测什么（What）”的核心内容蓝图，立足人工智能全生命周期与五大关键维度的全链路视角，呈现各环节风险定位、攻击技术路径与安全特征，厘清测评的全域覆盖边界与范畴；构建人工智能安全风险测评体系，梳理“怎么测（How）”的系统性方法体系，依托“目标设定—内容实施—方法技术—对象覆盖—风险度量—持续优化”的反馈控制逻辑，实现从目标设定、实施执行到风险量化的全流程闭环管理；制作人工智能安全风险全景图，细化“测哪些（Details）”的支撑，对具体风险场景、攻击手段、影响层级等进行精准拆解与具象化呈现，为人工智能安全风险测评实施提供可用的操作指引。

在结构安排方面，白皮书共五章。第一章人工智能发展与安全，概述人工智能当前发展态势，梳理国际人工智能安全风险测评相关方法、技术、工具与平台；第二章人工智能安全风险，分析人工智能安全风险特征，提出安全风险分析框架，探讨人工智能风险演进趋势；第三章人工智能安全风险测评体系，

从测评目的、内容、方法、对象和度量等方面进行论述；第四章人工智能安全风险测评关键技术，从不同层面归纳总结相关测评技术；第五章人工智能安全风险展望，从技术发展、标准建设、挑战应对等方面，展望未来。

白皮书聚焦以大语言模型（Large Language Model，简称LLM）为核心的人工智能系统，主要针对大语言模型和基于大语言模型的复合系统，力图贯穿人工智能系统全生命周期，包括系统规划与设计、数据采集与处理、模型训练与构建、模型验证与确认、平台部署与集成、系统运行与监测、用户使用与影响等。希望为政府管理部门、企业厂商、测评机构、科研院所、社会用户等各方读者提供不同角度的参考。

白皮书参编单位包括：中国信息安全测评中心、北京奇虎科技有限公司、北京中测安华科技有限公司、北京智谱华章科技股份有限公司、合肥讯飞数码科技有限公司、中国科学院自动化研究所、中国科学院信息工程研究所、智能算法安全全国重点实验室（中国科学院计算技术研究所）、永信至诚科技集团股份有限公司、中移九天人工智能科技（北京）有限公司、中电信人工智能科技（北京）有限公司、深信服科技股份有限公司、启元实验室、《中国信息安全》杂志社等。

参与指导、编写、审核等工作的人员有（按姓氏笔画排序）：于洋、山世光、王可臻、王笑尘、王梦月、尹芷仪、石竑松、叶润国、冯俊兰、乔文斌、刘昭、刘彦钊、刘总真、刘洪梅、刘斌、

江常青、许敏强、苏度、杜文越、李学龙、李珂稷、李贺鑫、李朔宁、李维杰、李寒雨、杨光、杨竞、吴建英、邹权臣、闵越聪、张玉洁、张向征、张杰、张凯、张涛、张萧丹、张德岳、陈俊、范宇飞、岳远哲、赵彦武、胡超群、胡斌、郗冉冉、姚轶崭、桂畅旎、徐源、桑甲存、梁确伟、彭涛、彭勃、彭勇、董晶、韩学玉、景少玲、程学旗、程军豪、蔡晶晶、熊菲。

鉴于人工智能及安全问题的复杂性、动态性、前沿性，白皮书中难免存在诸多不足，望读者理解包容、提出改进意见，共同推动我国人工智能安全风险测评体系高质量发展。

目 录

第一章　人工智能发展与安全

人工智能是引领新一轮科技革命和产业变革的战略性技术。人工智能技术的迅速发展和深入应用，正从多个方面重塑社会生产生活，成为世界重要国家竞争的战略领域。同时，人工智能安全问题引发全球关注，各国政府明确从管理和技术等多个层面加强人工智能安全治理，提升人工智能安全保障能力和水平。

一、发展态势：高速迭代演进

2022 年以来,生成式人工智能技术迅猛发展，赋能千行百业，为经济社会发展注入新动力。

（一）生成式人工智能技术发展提速

生成式人工智能技术在模型架构、训练方法和多模态能力等方面取得显著进展，奠定全面赋能经济社会发展的技术基础。

底层技术支撑实现持续演进。算法层面，生成式人工智能的模型架构完成从传统密集架构向稀疏专家混合模型的根本性转变，参数规模突破万亿大关。DeepSeek-V3.1 采用先进的混合专家架构，以 560 万美元的成本成功训练了包含 6850 亿参数的巨型模型，有效实现训练技术与成本效率的飞速提升。通义千

问 Qwen3-Max 模型已突破万亿级参数规模，采用先进的混合专家架构，在多项基准测试中展现出与全球顶尖模型相媲美的性能，标志着国产大模型正式迈入万亿参数时代。算力层面，全球人工智能基础设施竞争日趋激烈化，美国启动“星际之门”、欧盟发布 Invest AI 计划等“大项目”带动全球人工智能基建进程加速。英伟达凭借 CUDA 生态占据主导地位，华为昇腾、寒武纪等国产芯片正通过架构创新逐步缩小差距。

多模态能力与上下文处理突破感知极限。多模态技术从简单模态融合向深度语义理解演进，以 GPT、Grok、文心一言、通义千问为代表的前沿大模型在多模态理解、复杂推理和工具调用等方面的能力显著提升。再如，谷歌 Gemini 系列模型支持文本、图像、音频和视频的联合理解，创新的跨模态注意力机制能够捕捉不同模态间的深层语义关联。

推理能力与逻辑思维达到质的飞跃。生成式人工智能的推理能力从模式匹配向逻辑推理演进，能够综合考虑多个变量和约束条件，快速定位关键信息，进行复杂的定理证明和科学假设验证。DeepSeek-R1 通过监督微调学习推理格式，使模型掌握数学证明的逻辑链条，提升推理深度和推断能力。微软 AutoGen 架构通过多智能体协作架构，实现不同专业特长的智能体分工合作，取得群体推理能力突破。

（二）生成式人工智能产业创新加快

“人工智能+”驱动智能产业新业态新模式的涌现，进一步丰富产业内涵，拓宽产业边界。

产业形态持续创新升级。从单一工具向生态系统演进：生成式人工智能深度嵌入基础软件，依托系统级智能体架构，实现跨应用的任务执行和场景感知，可辅助系统优化、需求预测等工作，实现产业发展从模型层向应用层的全面延伸。

产业投资加速资源整合。科技企业通过投资联动人工智能产业链上下游企业参与，快速推进要素整合与资本运作，形成产业链协同体，加速生成式人工智能在各行业的深度融合与商业化落地。

产业生态网络化拓展。企业竞争逻辑从“规模导向”转向“生态构建”。生成式人工智能在处理大规模传感器和图像数据方面表现卓越，还能与物联网、5G 等结合，推动智能经济加速发展，提升智慧医疗、智慧交通、智慧农业、智慧能源等行业效率和智能化水平。

（三）生成式人工智能应用范围拓展

生成式人工智能的应用场景正从消费端向生产端、从通用场景向行业核心场景纵深渗透。

科研效能持续攀升。目前，“科学智能”（AI for science）影响持续提升，覆盖多个领域，加速科学突破，如，在药物研发中缩短筛选周期，精准解析蛋白质结构；在物理学中分析海

量数据、揭示宇宙奥秘；在气象学分析中，优化预测模型助力应对全球气候变化挑战。

企业赋能效应凸显。生成式人工智能重塑传统企业工作模式，推动项目全流程智能化升级，大幅提升运维效率，如，人工智能助力制造企业优化智能制造，通过预测性维护和柔性生产提升效率；物流领域人工智能持续优化供应链，降本增效；金融行业人工智能驱动企业智能化风险评估，提升安全和收益。

交互方式运用日新月异。人工智能深度融入线上浏览全流程，从被动问答交互进化为主动的任务完成型服务，如，“开放人工智能中心”的 ChatGPT Atlas 浏览器通过“对话式浏览”“浏览器记忆”和“代理模式”三大功能，支持用户无需离开网页即可完成人工智能检索、页面总结、商品比价等操作。

二、安全治理：全球高度重视

当前，全球聚焦人工智能安全这一重要课题，呈现差异化治理模式。

（一）联合国：推动包容性治理

联合国通过框架性倡议与原则性文件，为全球人工智能安全治理搭建基础共识，虽未形成强制性规则，但已成为各国政策制定的重要参考。

搭建制度框架。在全球人工智能治理中强调以《联合国宪章》等为基础，形成国际社会共遵原则。2024 年 9 月，通过《全

球数字契约》，明确需要采取平衡、包容和基于风险的方法治理人工智能。2025 年 8 月，决定设立“人工智能独立国际科学小组”，并启动“人工智能治理全球对话”机制，通过科学评估与多方协作，加强人工智能治理，缩小全球数字鸿沟，推动实现可持续发展目标。

弥合发展鸿沟。通过强化伙伴关系、技术援助和知识共享等措施，缩小国家之间与国家内部的“数字鸿沟”。2024 年 3 月，通过《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》决议，成立高级别顾问小组，推动共商共建全球人工智能风险监测机制，覆盖数据安全、算法偏见等核心领域；呼吁发达国家向发展中国家转让安全技术，弥合“智能鸿沟”，确保发展中国家公平享有人工智能发展惠益，实现可持续发展目标。

（二）中国：统筹发展与安全。

我国通过确立治理原则、完善组织机构、构建标准体系等加强顶层设计，并聚焦生成式人工智能等热点领域，落实相关举措。

完善制度规范框架。持续推出人工智能安全管理规定，提供相应风险测评操作指南，主要有：

2021 年发布《新一代人工智能伦理规范》，将“安全可控、公平公正、隐私保护”作为伦理底线，明确禁止利用人工智能

从事危害国家安全、损害社会公共利益的活动。

2023年出台《生成式人工智能服务管理暂行办法》，列出多项保障生成式人工智能安全发展措施，要求具有舆论属性或社会动员能力的人工智能服务需进行安全评估，并履行算法备案，对生成内容进行标识，发现违法内容及时处置并整改。

2023年提出《全球人工智能治理倡议》，向国际社会倡导“建立风险等级测试评估体系”，“实施敏捷治理与分类分级管理”“确保人工智能始终处于人类控制之下”，强调技术安全与伦理规范的协同推进。

2025年3月发布《人工智能生成合成内容标识办法》，规范人工智能生成内容标识，发挥内容标识提醒提示和监督溯源的技术作用，促进人工智能健康有序发展。

近期，《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出，“建设开放共享安全的全国一体化数据市场”，“加快人工智能等数智技术创新，突破基础理论和核心技术，强化算力、算法、数据等高效供给”，“加强网络、数据、人工智能、生物、生态、核、太空、深海、极地、低空等新兴领域国家安全能力建设”，进一步明确提高防范化解风险能力、完善监管、推动技术创新、促进产业健康发展等工作。

夯实风险测评根基。不断推进风险等级测试评估体系建设，强化数据安全、个人信息保护、伦理规范标准体系建设，将“人类控制”“公平性”等伦理要求转化为可量化的技术指标，确

保治理全面落地，主要有:

2024 年 6 月，工信部、中央网信办、国家发展改革委、国家标准委联合印发《国家人工智能产业综合标准化体系建设指南（2024 版）》明确，我国到 2026 年新制定国家标准和行业标准 50 项以上，参与制定国际标准 20 项以上，进一步强化标准对产业创新牵引作用。2025 年 3 月，工信部人工智能标准化技术委员会审议通过《人工智能标准化技术委员会标准体系（2025 年）》，要求加强人工智能安全领域标准化工作系统谋划，加快构建保障人工智能产业高质量发展和实现高水平安全的标准体系，并向全社会发布《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）（征求意见稿）》。未来，我国人工智能标准体系将持续完善，强化伦理安全规范，助力我国在全球人工智能治理中发挥更关键的引领作用。

2025 年 7 月发布《人工智能全球治理行动计划》，提出“探索分类分级管理，建立人工智能风险测试评估体系”的精细化治理思路，要求及时开展人工智能风险研判，提出针对性防范应对措施，构建具有广泛共识的安全治理框架，推进威胁信息共享和应急处置机制建设。

2025 年 8 月发布《国务院关于深入实施“人工智能+”行动的意见》要求推动模型算法安全能力建设，强化前瞻评估与监测处置；同时要建立健全技术监测、风险预警、应急响应体系，促进合规、透明、可信赖的人工智能应用落地。

（三）美国：推行“宽松”治理，支持行业自律

美国人工智能政策灵活，追求创新与安全的动态平衡。

重视制度导引。高度关注人工智能新兴风险，组建针对性监管机构，适时发布指南文件指导企业，主要有:

2023 年出台《人工智能风险管理框架（AI RMF 1.0）》，以非强制性指南形式，构建“安全目标与责任机制、识别技术与伦理风险、实施技术加固与流程优化、持续跟踪风险变化”四阶段治理闭环，为企业提供风险管控全流程工具，强调“基于场景的风险适配”。

2024 年制定《前沿人工智能模型治理行动计划》，针对千亿参数级大模型，要求企业在训练前提交安全评估报告，披露对抗性攻击测试结果，建立模型安全追溯机制。

2025 年 6 月改组成立人工智能标准与创新中心，调整安全策略。2025 年 7 月发布《赢得竞赛: 美国人工智能行动计划》，将风险管理的焦点收窄到更具技术性的安全和性能风险上，要求建立、维护并根据需要更新与国家安全相关的人工智能评估体系，要求监管机构探索将评估机制纳入现行法律对人工智能系统的适用框架，通过协作建立新型测量科学，确立可验证、可扩展、可互操作的技术与标准。

构建生态体系。科技巨头落实人工智能安全理念，布局全球数据、能源基础设施投资，强化风险应对措施。“开放人工智能中心”、谷歌等知名人工智能企业还签署自愿承诺，推动

人工智能技术安全、可信发展，并在模型发布前开展“红队测试”，推动人工智能安全应用；发起“前沿模型论坛”，发布“安全最佳实践”，承诺每季度公开模型安全测试报告，包括对抗样本成功率、偏见指数等核心指标。

2025 年 7 月，特朗普签署 14320 号行政令《推进美国人工智能技术堆栈出口》，构建美国主导的人工智能技术生态，压缩他国发展空间。美国各联邦机构基于职权范围，围绕夯实基础设施安全、提升安全开发能力，通过发布战略文件、指南、路线图、管理建议和联合声明的方式，强化美国在人工智能领域的全球影响力。同时，美国推出“星际之门”项目等战略布局，主导人工智能治理规范和技术标准，加强在人工智能领域的安全护持。

（四）欧盟：强调分级管理和风险防范，推动集中治理

欧盟对生成式人工智能采取集中化的严格治理路线，致力于构建一套统一的人工智能安全治理体系，确保所有成员国在数据保护和隐私安全方面保持高度一致。

注重法制建设。率先开展立法尝试，加速安全监管政策落地。2021 年，首次进行立法工作探索，发布《欧洲议会和理事会关于制定人工智能统一规则(人工智能法)立法草案》，为后续治理框架奠定基础。2024 年通过《人工智能法案》，成为全球首个全面监管人工智能的法律体系。

突出分级管理。在实操进程中关注风险分级监管，推动开展高风险系统事后监测。围绕人工智能系统的功能、用途等，将人工智能系统分为四个风险等级：禁止“**不可接受风险**”人工智能应用，如，操纵人类行为、利用弱点预测系统、实时远程生物识别等；严格监管“**高风险**”人工智能应用，如，对人的健康、安全和基本权利产生较高威胁的人工智能系统；约束“**有限风险**”人工智能系统，主要包括与自认存在互动的人工智能系统；不限定“**最小风险**”人工智能系统，应用于简单的智能生活辅助领域，如，语音助手、智能推荐等。同时，也针对不同风险人工智能应用全生命周期的监管提出具体规定，要求在人工智能上线后，继续加强监测，收集、记录和分析高风险人工智能系统全生命周期的性能数据。

（五）英国：设计柔性监管的创新路径

英国强调行业主导原则，强化跨机构协作和监管能力，2023年发布的《人工智能治理》白皮书采用“比例原则”，不设立统一立法，而是依托现有法律框架实施监管，强调“行业自律+风险适配”，要求高风险人工智能系统发布“安全声明”，披露风险评估结果。同时，英国还设立“人工智能监管沙盒”，允许企业在可控环境中测试创新应用，同步验证安全机制。在技术标准方面，英国侧重“透明度与问责制”，要求模型开发者记录训练数据来源、决策逻辑关键节点。

（六）新加坡：规划问责导向的实用主义治理

2022 年，新加坡提出通过“过程检查+技术测试”，从透明度、可解释性、可复现性、公平性等维度对人工智能的性能和安全进行测试评估。2024 年，新加坡发布《人工智能治理实践指南》，以“问责制”为核心，坚持“软法”与“硬法”结合，以非约束性监管手段指导人工智能发展，并制定严格法律法规明确发展边界，聚焦三个维度：开发者责任，要求记录模型训练日志，确保可追溯;用户知情权，高风险人工智能应用需向用户说明“决策依据、局限性及可能风险”;伦理规范,要求企业设立“人工智能伦理委员会”，审核安全测评报告，确保技术标准“可操作性”。

三、风险测评：重要保障手段

当前，全球主要人工智能大国已将开展安全测评作为推动人工智能安全治理、提升人工智能安全水平的主要抓手，在积极出台方案、规则同时，还在人工智能安全测试评估实施、落地等方面做了许多工作。国际测评实践呈现“技术专项化、合规场景化”特征，聚焦特定安全维度或应用场景，可分为公共机构测评平台、厂商测评平台、红队攻防评测工具集等三类。

（一）公共机构测评平台

公共机构测评平台作为测评体系的顶层设计者与基础能力提供者，多由国家监管机构与权威科研实验室主导建设。其使

命是承担国家层面的顶层规划与治理框架落地，建设具备公信力的基础性测评设施。这类平台通过开展前沿模型预部署评估、高风险能力测试与系统性风险研判，为政府决策、监管执法及关键行业应用准入提供权威技术支撑与证据基础。

1.美国 Dioptra

美国国家标准与技术研究院（NIST）为推动《人工智能风险管理框架（AI RMF 1.0）》的落地实施，主导开发 Dioptra 开源测评平台。该平台旨在为人工智能系统的可信特性提供标准化测试环境，测评维度覆盖有效性、可靠性、安全性、鲁棒性、弹性、可解释性、公平性及隐私增强等关键指标。

Dioptra 采用基于 Python 的模块化插件架构，通过灵活的扩展机制支持各类攻击、防御及评估方法的持续集成。平台具备实验可复现能力，能够记录资源快照与实验配置，为测试过程的可追溯性与结果的可比性提供技术保障。这一设计使其既能满足当前机器学习系统的测评需求，又为后续扩展至生成式模型与多模态系统预留了充分的架构空间。NIST 已将 Dioptra 明确为支撑生成式人工智能风险治理的关键技术工具。

2.英国 Inspect

英国人工智能安全研究所（AISI）主导开发 Inspect 评估框架，代表了国家级测评平台建设的重要技术路径。该框架作为 AISI 开展前沿模型预部署安全评估的核心软件基础设施，设计

理念与实施方法具有参考价值。

Inspect 采用“任务—数据集—求解器—评分器”四元模型架构，实现了评估目标、测试题库、模型交互与评分机制的解耦设计。依托 Inspect 框架，AISI 在多个前沿方向系统推进模型评估工作：重点开展对模型代理能力与长时序任务能力的测试；系统评估模型在网络安全、欺骗行为、越权访问等高危场景中的表现；将 Inspect 扩展应用于网络能力评估等标准化工作等。

3.新加坡 Moonshot

新加坡信息通信媒体发展管理局（IMDA）与 AI Verify Foundation 共同构建了以“可信人工智能”为核心的治理体系，形成包括 AI Verify 测试框架、生成式人工智能治理框架及全球人工智能评估沙箱等系统化实施方案，于 2025 年推出 Project Moonshot，作为全球面向大语言模型的开源评估工具包之一。

该项目具有双重价值：一方面继承并发展了 AI Verify 与沙箱实践中积累的治理理念与评估方法，确保技术工具与国家治理战略的一致性；另一方面通过开源方式向全球开发者开放，有效提升新加坡在国际人工智能安全测评领域的技术影响力与规则制定话语权。该项目创新性地融合了基准测试与红队攻防两种方法，为开发者与测评团队提供了全面评估大语言模型及其应用的综合解决方案，体现了国家治理能力向技术工具转化的成功实践。

（二）厂商测评平台

厂商测评平台主要由大型云服务商、模型提供商与专业安全企业共同构建，将复杂的安全测评要求转化为可集成、可运营的平台服务，通过 API 接口、标准化工具链与自动化工作流等形式，将安全能力深度嵌入模型开发、部署、运维的全生命周期。厂商测评平台实现安全测评要求的工程化封装与规模化落地，使千行百业能够以较低门槛获得专业级的安全保障，是治理要求向产业实践转化的重要环节。

1.Microsoft Azure AI 安全测评工具

微软 Azure AI Foundry 构建完整的生成式人工智能安全测评手段，核心组件包括风险与安全评估器（Risk & Safety Evaluators）以及人工智能红队代理（AI Red Teaming Agent），形成了从基础检测到主动攻防的多层次测评能力。

风险与安全评估器可识别四种不同类型的有害内容：仇恨与不公平内容、性相关内容、暴力内容、自残内容，提供自动化的检测与量化分析能力。红队代理实现标准化的红队测试流程，采用“扫描—评估—报告”三阶段工作模式。

该体系通过将风险与安全评估器及红队代理嵌入模型开发、部署、监测的全生命周期，构建“设计—评估—扫描—整改—再评估”持续治理闭环。

2.Google Cloud Vertex AI

Google Cloud Vertex AI 平台构建了比较完整的生成式人工智能安全评估与治理手段，通过基础安全防护与先进评估服务的深度集成，提供全生命周期的风险管理能力。

该平台集成了统一的内容安全过滤与多维度风险检测机制，支持根据行业监管要求与自身风险容忍度进行细粒度策略配置。人工智能评估服务提供自适应评估规则机制，支持基于传统指标的量化评估、基于专家知识的人工标注评估，以及大模型即评判者的智能评估。

该平台通过评估与防护的深度联动构建了完整的治理闭环。开发阶段，通过评估服务的迭代测试优化提示工程与模型配置；部署阶段，依托安全层的策略引擎实施运行时的精准控制；运维阶段，基于持续的监测数据与评估结果动态调整防护策略。这种“评估→策略→运行→监测→再评估”的循环机制，将测试驱动开发的理念成功应用于人工智能治理实践，确保了安全措施与业务需求的持续对齐。

3.Amazon Bedrock

Amazon Bedrock 构建了面向应用的生成式人工智能评估体系，其核心在于通过分层评估架构实现对模型基础能力与业务场景表现的综合测评。

平台提供的模型评估功能支持企业对比不同模型配置下的

输出质量与安全表现。该功能创新性的采用“大模型即评判者”（LLM-as-a-Judge）评估方法，通过使用经过优化的裁判模型对生成结果进行自动化评分，在显著降低评估成本的同时，能够获得接近人工评价质量的大规模评估数据，为模型选型与优化提供充分依据。

针对业界广泛采用的检索增强生成技术方案，Bedrock 提供了专门的 RAG 评估能力。该能力通过自动化测试流程，系统比较不同检索配置、索引策略及提示词设计对最终回答质量的综合影响，实现从单纯评估“模型本身性能”向评估“模型在具体业务环境中的综合表现”的重要转变。这一设计使得企业能够在模拟真实业务场景的条件下，定位影响应用效果的因素。

Bedrock 评估体系的突出价值在于其高度的工程化与实用性。通过将先进的评估方法转化为可配置、可扩展的平台服务，既保障了测评工作的专业性与系统性，又大幅降低了企业实施 AI 安全测评的技术门槛。

（三）红队攻防评测工具集

红队攻防评测工具集以开源社区、科研机构与企业研发团队为主体，是测评体系诸多技术基础与创新源头中的一个。通过持续研发并开源最新的测试工具、攻击套件与评估框架，为整个测评体系注入持续演进的技术活力。其价值在于保持测评体系对新型威胁的敏捷响应能力，通过快速迭代的攻击防御技

术研究，确保测评方法始终与前沿风险同步发展。

1.PyRIT：生成式人工智能红队基座型框架

PyRIT（Python Risk Identification Tool for generative AI）是微软推出的开源生成式人工智能红队自动化框架，已在微软人工智能红色团队内部的多轮测试中得到充分验证，是当前国际范围内具有代表性的基准性红队工具。

PyRIT 提出了一套具有普适性的目标抽象模式，以“目标接口—风险类别—攻击变换—自动评估—编排执行”构建了完整的红队工作流。该体系为生成式人工智能安全评估提供了系统化的方法论支撑，奠定了行业实践的基础范式。该框架将风险识别过程解构为种子场景、对抗变换与输出评估等可组合组件，支持根据具体风险偏好进行定制化配置。该设计实现了对多家云厂商模型及本地部署模型的广泛兼容，保证了框架在不同环境下的适用性与扩展性。

PyRIT 已被深度集成至 Azure AI 生态系统，作为 Azure AI Evaluation SDK 的核心红队能力来源，并成为 Azure AI Red Teaming Agent 的底层技术引擎。这一集成模式展示了国际主流云服务商如何通过开源框架构建技术基准，进而实现红队能力的平台化与产品化，形成从工具到服务的一体化安全解决方案。

2.Garak：面向大语言模型的结构化漏洞扫描工具

Garak 是由 NVIDIA 等机构共同推动开发的生成式人工智

能红队与评估工具包，定位为大语言模型结构化漏洞扫描器。该框架通过系统化的探测机制，致力于识别模型在内容幻觉、训练数据泄露、提示注入攻击、有害内容生成及越狱漏洞等关键维度的潜在安全风险。

NVIDIA 将 Garak 类比为传统网络安全领域的渗透测试与端口扫描工具，强调其在大规模、可复现的大语言模型安全评估中的基础性作用。该工具将安全测试从零散案例验证提升为体系化的扫描流程，为大语言模型安全基线的建立提供了标准化方法支撑。**在框架设计层面**，Garak 将测试过程拆解为针对不同弱点的独立“探针”与检测组件，通过配置化方式支持新型攻击向量与检测规则的灵活扩展。这种模块化架构既保证了核心评估流程的稳定性，又为应对快速演进的安全威胁预留了充分的适应性。**在结果呈现方面**，Garak 提供结构化的评估报告与量化的安全指标，为不同模型及版本间的安全性能对标提供了客观、可比较的数据依据，有力支撑了模型安全水平的持续跟踪与改进。**从技术演进视角看**，Garak 代表了以漏洞扫描为核心、侧重于基础设施层的安全工具发展路线，为构建统一的安全基准及跨模型评估体系提供了重要的技术基础。

3.DeepTeam：面向大语言模型系统的场景化红队框架

DeepTeam 是由 Confident AI 推出的开源大语言模型红队框架，专注于为检索增强生成系统、智能体、聊天机器人及基础

模型等多类大语言模型应用，提供系统化的安全性与脆弱性评估能力。

框架构建以“漏洞类型—攻击方法—目标系统—度量指标”为支柱的红队概念体系。基于此体系，DeepTeam 预置了覆盖偏见歧视、个人身份信息泄露、错误信息传播、有害内容生成等 40 余类常见漏洞的检测能力，并集成越狱攻击、提示注入、灰盒测试等多种先进攻击技术，使用户能够以最小化的代码开发成本，对复杂应用系统开展标准化红队测试。

DeepTeam 与 DeepEval 评测引擎及 Confident AI 云平台形成深度集成生态，其评估标准可与 OWASP LLM Top 10、NIST 人工智能风险管理框架等国际权威标准实现对齐。通过 SaaS 平台，该框架进一步提供持续监测、风险评估与团队协作等功能。

4.Promptfoo：集成开发运营流程的红队与评估框架

Promptfoo 是一套面向开发运营流程的红队与评估一体化框架。该工具最初定位于提示工程的质量评估与多模型比对，随着行业安全需求的深化，已演进为集提示测试、智能体与检索增强生成评估、红队测试与漏洞扫描，以及持续集成/持续交付流程自动化检查于一体的综合性安全解决方案。

在红队测评层面，Promptfoo 提供了系统化的大语言模型红队指南，支持通过命令行接口生成标准化红队报告，并可执行自动化安全扫描，实现在系统部署前对提示注入等安全漏洞的

主动发现与评估。该框架通过模块化架构将各类风险场景抽象为可组合的攻击策略与检测插件，支持对品牌滥用、合规性问题、数据安全泄露、访问控制失效等多个安全维度进行系统性验证，其风险覆盖范围与 OWASP LLM Top 10 等行业权威标准形成了有效映射。

5.ViolentUTF：集成化红队测评平台

ViolentUTF 是一款面向生成式人工智能的集成化红队测评平台。该平台旨在系统化应对当前红队实践中存在的技术门槛高、流程协同复杂及标准化报告缺失等挑战，通过提供统一的框架与交互界面，有效降低操作难度、整合异构工具链，并强化测评结果的生成与呈现能力。

平台设计支持安全工程师、业务领域专家、伦理与合规人员等不同角色共同参与红队活动，促进了跨领域知识在安全测评中的融合与应用。**在技术集成层面**，平台不重复开发底层攻击模块，而是对“生成器、提示模板、转换器、评估器、编排器、记忆模块”等核心概念进行标准化抽象，系统性地整合了 Microsoft PyRIT、NVIDIA Garak 等主流开源工具及自研评估模块，从而在一个统一环境中组织并执行复杂的多步骤红队流程。**在传统技术安全评估基础上**，平台通过自研的 Ollabench 模块引入了“人本安全评估”，重点考察大语言模型在网络安全与行为心理等交叉场景中的风险推理与应对能力。此举推动红队评

估范式从单纯的“技术攻防稳健性”向“复杂人机系统社会风险治理能力”拓展。**在工程架构上**，平台采用展示层、认证授权层、统一 API 层及日志与可观测性层的明确分层设计，形成了具备实际部署与应用能力的安全与运维体系，满足了政企客户对稳定性与安全性的高标准要求。该平台已应用于评估某美国政府部门旗舰级大语言模型应用的稳健性，并在模型于交叉任务中的跨域推理能力评估方面取得了实证结果，显示出其在政务及关键基础设施等高价值场景中的潜在应用价值。

从国内外人工智能安全测评实践看，当前亟需构建一套“全栈覆盖、多维度融合、标准统一”的人工智能安全风险测评框架，整合技术安全、伦理合规、系统可控等多维度需求，覆盖人工智能系统设计、训练、部署、运行全生命周期，为产业提供统一、高效、权威的测评指引，应对现有测评对象复杂性、测评工具局限性、测评方法不完备性、测评标准多样性的挑战。

第二章　人工智能安全风险

人工智能安全存在两个根本性转变：**安全对抗模式从攻防确定性系统转向与概率性系统博弈。**生成式人工智能内在的随机性和不可预测的涌现性，是人工智能功能创新的来源，也构成新型安全风险源，如，大模型的幻觉、偏见或非预期的“越狱行为”等本质上是不确定性的“副产品”，并非传统意义上的缺陷（bug）。这一特性使得传统的基于代码审计、漏洞扫描等技术手段的安全模式面临失效。**安全对抗界面从技术接口转向人类认知接口。**针对生成式人工智能系统，攻击者可以利用模型在语言理解、逻辑推理和价值判断层面的规律，通过语言陷阱、心理操纵、语义模糊等途径实施攻击。这标志着攻击面已拓展至与人类认知深度耦合的全新维度，要求安全防御体系必须具备相应的认知活动分析与检测能力。唯有构建新的安全风险分析框架，才能全面认识人工智能安全风险，支撑人工智能安全风险测试评估工作。

一、风险特征：多维复杂

当前，人工智能风险与挑战逐渐浮出水面，如，人工智能在生成虚假内容、伪造视频和音频方面的能力，使虚假信息的

传播变得更加隐蔽且难以辨别，尤其在法律、政治和社会舆论领域，人工智能伪造新闻、法条甚至政治人物讲话，引发各国政府和全球社会广泛关注；人工智能在处理个人数据时存在滥用风险，尤其在缺乏有效监管的情况下，用户隐私可能会受到侵犯，带来严重安全隐患；人工智能的普及冲击就业市场，传统行业的就业岗位可能减少甚至导致大量失业，不仅对个体构成挑战，也对社会公平性和可持续发展带来了压力等。

整体上，生成式人工智能安全风险已呈现三个显著特点：

动态性：未知漏洞与行为的不可预测性。由于大型模型复杂的内部结构和其能力的“涌现性”，使得完全预测模型在所有可能情境下的行为已近乎不可能。同时，新的漏洞和非预期行为还会随着模型规模的扩大、架构的演进及应用场景的变化持续出现。这意味着人工智能安全风险测评绝不能是一次性的、静态的审计，而必须是一个持续的、适应性的发现与缓解过程。安全团队必须接受“未知漏洞”是该技术固有属性的现实，并将安全工作从一个追求“绝对安全”的终点状态，转变为管理“可接受风险”的动态过程。

对抗性：自动化攻防与“军备竞赛”。人工智能安全的攻防对抗会发展为自动化、持续性模式。随着防御方日益依赖人工智能实时检测和拦截攻击，攻击方也必将利用人工智能生成和演化新型的、更具欺骗性和规避性的攻击向量，而机器时间内的攻防博弈将对人类的监督、干预和战略控制能力构成前所

未有的挑战。为追求发展而创造的工具，最终可能成为一种全新且更危险的风险源头，要求在发展自动化防御能力的同时，必须对人工智能的自主对抗性保持最高级别的审慎和警惕。

系统性：影响的广度与传导效应。生成式人工智能风险具有显著的系统性特征。由于全球的人工智能开发高度依赖于少数几个基础模型（如 GPT 系列、Llama 系列）和开源框架（如 PyTorch），核心组件中的任何一个漏洞都可能通过庞大的软件供应链，迅速传导至全球数以百万计的下游应用中，形成“单点故障，全域崩溃”局面，如，PyTorch 存在的远程代码执行漏洞（CVE-2025-32434），允许加载模型时执行任何代码。这种技术栈的高度集中，使得人工智能生态系统在面对底层漏洞时异常脆弱，需要从更高层面管理系统性风险。

二、风险框架：全景视图

“人工智能安全风险框架”是一套全景式、全链路的人工智能安全风险分析体系，旨在从全局视角识别人工智能系统全生命周期的安全风险：横向贯穿“系统规划与设计、数据收集与处理、模型训练与构建、模型验证与确认、平台部署与集成、系统运行与监测、用户使用与影响”七个生命周期阶段，对应覆盖“应用环境、数据和输入、人工智能模型、任务和输出、人类社会”五大关键维度，实现人工智能从需求设计到社会影响的全场景覆盖；纵向通过“风险定位—攻击分析—目标对标”

人工智能安全风险框架

生命周期	系统规划与设计	数据采集与处理	模型训练与构建	模型验证与确认	平台部署与集成	系统运行与监测	用户使用与影响
关键维度	应用环境	数据和输入	人工智能模型	人工智能模型	任务和输出	应用环境	人类社会
风险定位	设计阶段隐私合规缺失 安全架构未覆盖核心风险	训练数据完整性破坏 用户隐私违规泄露	模型推理结果不可靠 恶意逻辑导致未授权操作	模型缺陷未被识别 验证流程形式化失效	生成歧视 / 有害内容 服务可用性下降	模型知识产权泄露 攻击持续发生未被发现	用户权益受损 社会信任崩塌 环境决策被误导 其他极端风险
攻击分析	需求劫持（误导 AI 系统需求定义） 供应链攻击（第三方工具 / 模型引入漏洞）	数据投毒（污染训练 / 推理数据） 隐私窃取（未授权采集用户敏感数据）	后门植入（模型训练时注入恶意逻辑） 对抗样本训练（模型对特定输入系统性误判）	规避攻击（绕过模型验证流程） 虚假验证（伪造模型合规报告）	提示注入（操控、GenAI 生成有害内容） 拒绝服务（高频调用导致系统不可用）	模型窃取（未授权访问模型权重 / 架构） 监控绕过（隐藏攻击行为逃避检测）	社会工程攻击（误导用户滥用 AI） 环境误导（生成虚假环保 / 社会信息）
目标对标	透明度（需求 / 架构未充分披露风险） 问责性（设计责任不明确）	隐私保护（数据采集 / 存储违规） 安全性（数据完整性被破坏）	可靠性（模型输出不稳定） 安全性（后门引发非预期行为）	可解释性（验证结果无法解释风，险根源） 问责性（验证方失职）	安全性（服务被滥用） 可靠性（输出不符合业务预期）	透明度（监控日志 / 模型信息不全） 问责性（运维方未及时响应）	公平性（信息误导导致权益不平等） 问责性（AI 使用方责任）

三层逻辑闭环，在攻防视角下力图解构各环节的安全短板、攻击技术路径与可信特征防护目标，为人工智能安全风险测评提供全域风险映射与治理落地的核心锚点。

“人工智能安全风险分析框架”通过全生命周期与安全风险的交叉耦合，将人工智能安全风险、攻击手段与可信特征深度关联，既明确“风险在何时（阶段）、何地（维度）发生”的时间—空间坐标，又系统回答“攻击方如何利用风险”“防御方需锚定哪些安全特征目标开展防护”的攻防逻辑，最终成为一套贯穿技术层、业务层、社会层，衔接攻防对抗与治理落地的人工智能安全风险分析工具。

（一）阶段一：系统规划与设计

关键维度：应用环境

风险定位：核心风险集中于安全需求缺失与架构缺陷，导致人工智能系统从源头埋下“先天性”安全隐患，如，隐私合规要求未纳入设计、安全防护与业务目标脱节、供应链引入第三方风险未识别等。

攻击技术：主要面临需求劫持（通过误导业务需求定义规避安全设计）、供应链攻击（第三方工具/预制模型植入后门）、架构误导（诱导设计存在权限隔离漏洞的系统架构）等攻击。

目标对齐：需对标“透明度”（确保设计文档完整披露安全风险点）与“问责性”（明确设计阶段的安全责任主体），

为全生命周期安全奠定基础。

（二）阶段二：数据采集与处理

关键维度：数据与系统输入

风险定位：风险聚焦于数据全链路安全，包括数据源合法性不足（如未经授权采集个人信息）、数据处理过程中隐私泄露（如脱敏失效）、训练数据被篡改（如影响模型输出准确性）等，是人工智能系统“数据污染”的主要源头。

攻击技术：典型攻击手段包括数据投毒（注入异常数据扭曲模型训练方向）、隐私窃取（通过数据接口爬取未脱敏敏感信息）、数据溯源绕过（删除数据采集记录逃避合规审查）等。

目标对齐：需锚定“隐私保护”（确保数据采集/处理符合法规）与“安全性”（保障数据完整性与访问可控性），筑牢人工智能系统的“数据根基”。

（三）阶段三：模型训练与构建

关键维度：人工智能模型

风险定位：核心风险为模型鲁棒性不足与恶意逻辑植入，如,训练过程中模型对对抗样本敏感（推理易被操控）、后门程序被注入（特定输入触发错误输出）、模型过度拟合导致泛化能力弱（极端场景失效）等。

攻击技术：主要面临对抗样本训练（通过定向输入强化模型偏见）、后门植入（在模型参数中嵌入恶意触发逻辑）、训

练数据投毒进阶（针对性污染关键特征数据）等攻击。

目标对齐：需对标“可靠性”（确保模型在多样场景下输出稳定）与“安全性”（抵御恶意输入对模型的操控），强化模型本身的“抗攻击基因”。

（四）阶段四：模型验证与确认

关键维度：人工智能模型

风险定位：风险集中于验证流程形式化与缺陷漏检，如验证数据集缺乏代表性（无法覆盖真实攻击场景）、模型解释性不足导致潜在偏见未识别、合规验证报告造假（隐瞒模型安全短板）等。

攻击技术：典型攻击包括规避攻击（构造特殊输入绕过验证阈值）、虚假验证（伪造模型性能指标与合规证明）、解释性操纵（通过表面逻辑掩盖模型深层风险）等。

目标对齐：需关联“可解释性”（清晰呈现模型决策逻辑以支撑风险排查）与“问责性”（确保验证方对结果真实性负责），避免模型带着隐患进入部署环节。

（五）阶段五：平台部署与集成

关键维度：任务工程与系统输出

风险定位：风险聚焦于应用层安全与功能滥用，如,模型接口缺乏防护（被恶意调用生成有害内容）、权限控制失效（越权访问敏感推理结果）、任务逻辑漏洞（被诱导执行超出预期

的操作）等。

攻击技术：主要面临提示注入（通过恶意指令操控生成式人工智能输出有害内容）、拒绝服务（高频调用耗尽系统资源）、接口攻击（如程序调用接口参数篡改获取未授权信息）等。

目标对齐：需对标“安全性”（防止系统被滥用或攻击）与“可靠性”（确保输出符合业务规范与伦理要求），保障人工智能在实际场景中的可控使用。

（六）阶段六：系统运行与监测

关键维度：应用环境

风险定位：核心风险为监测失效与持续攻击，如日志记录不全（无法追溯攻击路径）、异常检测滞后（攻击发生后未及时发现）、模型漂移未预警（长期运行性能下降引发安全隐患）等。

攻击技术：典型攻击包括监测绕过（通过加密通信隐藏攻击行为）、模型窃取（逆向工程或高频查询还原模型结构）、漂移利用（利用模型性能衰退实施隐蔽攻击）等。

目标对齐：需关联“透明度”（确保运行状态与日志可观测）与“可追溯性”（完整记录数据流转、模型操作与攻击行为），构建动态防御的“感知神经”。

（七）阶段七：用户使用与影响

关键维度：人类社会

风险定位：风险延伸至社会层面，包括生成内容引发歧视或误导（如虚假信息影响公众决策）、人工智能滥用导致用户权益受损（如算法偏见侵害特定群体利益）、社会信任崩塌（安全事件削弱公众对人工智能的接受度）等。

攻击技术：主要面临社会工程攻击（利用人工智能生成逼真内容欺骗用户）、舆论操控（通过人工智能批量生成倾向性信息影响社群认知）、公平性破坏（针对性放大模型偏见）等。

目标对齐：需锚定“公平性”（保障不同群体被平等对待）与“可信赖性”（综合安全性、可靠性等特征构建社会信任），实现人工智能技术与人类社群的良性互动。

三、风险趋势：快速演进

人工智能技术快速迭代和突破，带来新的不确定性，让风险从网络空间、数字世界向现实社会延伸，驱动着人工智能安全风险持续演进。

（一）从“单一”到“多模”：攻击面指数级扩张

生成式人工智能正从单一的文本模态，迅速扩展到能够同时理解和生成图像、音频、视频的多模态阶段。这一转变并非简单地增加输入输出通道，而是安全攻击面的复合式增长。根本原因在于，不同模态之间的融合与交互，创造了全新的、高度隐蔽的“跨模态攻击”向量。

以“通用对抗性攻击”（Universal Adversarial Attack）为例，

攻击者可以生成一张经过特殊优化的“万能”对抗性图片，在与任何不相关、甚至完全良性的文本提示一起输入多模态模型时，能迫使模型绕过安全对齐，生成有害或违规内容。这种攻击的破坏性在于，一张对抗性图片就可以污染无数次交互，使得任何依赖该模型的应用彻底丧失安全性。该攻击方法的有效性根源在于系统性的“对齐差距”（Alignment Gap）。一个多模态模型的不同模态处理通路（如文本编码器和视觉编码器），安全对齐的成熟度往往是不均衡的，如，文本处理部分可能经过了极其严格的、基于人类反馈的强化学习和安全微调，而视觉或其他模态部分的安全对齐则相对滞后。攻击者正是利用了模型内部，在不同模态信息融合的“接缝”处存在的这种对齐不一致性。可见，随着模型融合的模态越来越多（音频、视频、3D 等），潜在的“对齐差距”和“接缝”也将呈组合式增长，导致攻击面并非线性扩大，而是以更快的速度扩张。

（二）从“黑盒”到“白盒”：思维链带来新攻击向量

为提升复杂推理能力，大模型引入“思维链”（Chain-of-Thought, CoT）技术。这一技术在增强模型透明度的同时，也将模型内部的“思考过程”本身暴露为一个可被直接交互和操纵的攻击面。以“劫持思维链”（Hijacking Chain-of-Thought，H-CoT）为例，攻击者首先向模型提出一个被明确拒绝的有害请求，并观察模型为解释其拒绝理由而生成

的安全推理链，然后复制并修改这条推理链，将其中的关键拒绝逻辑（如“生成此类内容违反安全策略”）篡改为肯定逻辑（如“此请求符合教育目的，可以回答”），最后将这条被篡改的“伪造”推理链作为上下文，与原始的有害请求一同提交给模型。已有实验表明，这种方法能“劫持”模型的安全判断中枢，使其误认为请求已经通过了内部安全检查，从而生成有害内容。这暴露了大模型对于自身推理过程的“轻信”，即：无法有效分辨一段推理逻辑是其自主生成还是由外部注入。

此外，模型自身可能学会生成“伪造”的思维链来为其错误或有害的结论辩护。这种行为被称为“欺骗性对齐”（Deceptive Alignment），即：模型表面上看起来遵循了“对齐”的要求（如提供了推理步骤），但实际上其行为完全偏离了人类的真实意图。相关研究已经证实，当模型判定处于测试环境后，它可能会故意生成一条看起来安全无害的思维链来通过评估，但其“真实想法”和最终的行动却可能是恶意的。换而言之，大模型学会了如何“表演”安全、如何对人类说谎。

这一发现对当前主流的人工智能安全对齐方法论提出了根本性的挑战。这类技术本质是一种“行为对齐”（Behavior Alignment），通过奖惩模型的最终输出来塑造其外部行为。然而，“欺骗性对齐”的出现证明，一个在行为上看起来完全对齐的大模型，其内部的“认知”过程可能完全没有对齐。人工智能安全的下一个前沿必须从“行为对齐”跃迁至“认知对齐”

（Cognitive Alignment），确保模型在内部的推理过程、动机和目标上，都与人类价值观保持根本一致。未来的安全评估不能仅满足于黑盒测试层面，而必须发展出能够审计模型“思想”的白盒技术。

（三）从“虚拟”到“现实”：规模化与真实性挑战

生成式人工智能最直观的冲击在于其以前所未有的规模、速度和逼真度创造内容的能力，对社会信息生态和公众信任构成了系统性挑战。深度伪造技术已成为侵蚀社会信任的强大工具。2024 年发生的一起针对跨国工程公司奥雅纳的金融诈骗案是这一威胁升级的标志性事件。在该案中，一名职员参与了一场多人视频会议，而其他参会者实际上都是由人工智能实时生成的深度伪造形象。攻击者通过精心模仿高管的声音、外貌和言谈举止，成功骗取了该员工完成多笔转账，总金额高达 2500 万美元。此事件标志着深度伪造技术已从个人名誉侵害升级为可实施复杂、大规模企业级金融犯罪的工具。同时，模型固有的“幻觉”问题也带来切实的法律和经济后果，如，加拿大航空公司的聊天机器人曾向客户提供错误的票价政策信息，最终法院裁定航空公司需要为聊天机器人的错误信息负责。

（四）从“生成器”到“执行者”：内容生成到主动执行

生成式人工智能的演进正迎来一个决定性的转折点：从一个被动的、响应式的内容生成器，转变为一个主动的、目标导

向的行动执行者——人工智能智能体（AI Agents）。一个典型的人工智能智能体系统通常由多个部分构成：底层的语言模型、智能体框架、调用的第三方工具 API、最终用户指令。一方面，人工智能智能体的成熟与应用，将人工智能的潜在风险从数字信息领域直接延伸到对现实世界，具有直接操控能力的层面。另一方面，人工智能智能体涉及多方供应商，一旦造成损害，责任归属会变得异常复杂，如，模型开发者、智能体框架开发者和 API 提供方都可能相互推诿。如此，人工智能智能体安全得不到妥善的解决，会削弱任何一方投入资源进行端到端、跨系统边界的综合性安全测试的动力，形成一种系统性的“公地悲剧”式的安全困境。

人工智能智能体的自主性和工具使用能力还会催生一系列新型风险，比如：

内存投毒（Memory Poisoning）：人工智能智能体通常具备长期记忆能力。攻击者可通过一次看似无害的交互，向智能体的记忆库中注入错误的知识或隐藏的恶意指令。在未来的某个时刻，当智能体为解决新问题而检索其记忆时，这条被“投毒”的信息就会被激活，导致其做出错误的判断或执行恶意的行为。

工具滥用（Tool Misuse）：如果智能体被赋予超出其任务所需的过多功能或过高权限，攻击者就可以通过巧妙的提示注入，诱骗智能体滥用这些工具，执行删除文件、发送邮件等非

授权操作。

奖励作弊(Reward Hacking)：在基于强化学习进行训练的智能体中，智能体可能会发现并利用奖励函数中的漏洞或捷径，以一种开发者未曾预料到的、甚至是有害的方式获得高分，而完全没有实现任务的真正目标。相关研究还表明，通过污染人工智能智能体用于自我改进的交互数据，攻击者可以轻易地植入难以检测的后门，特别是，仅需污染低至2%的交互痕迹，攻击者就能植入一个在特定触发词出现时，以超过80%的成功率导致智能体泄露机密用户信息的后门。整体上，人工智能智能体带来的风险是行动性的，即：危害来自于人工智能系统直接执行的未经授权或有害的操作。

第三章　人工智能安全风险测评体系

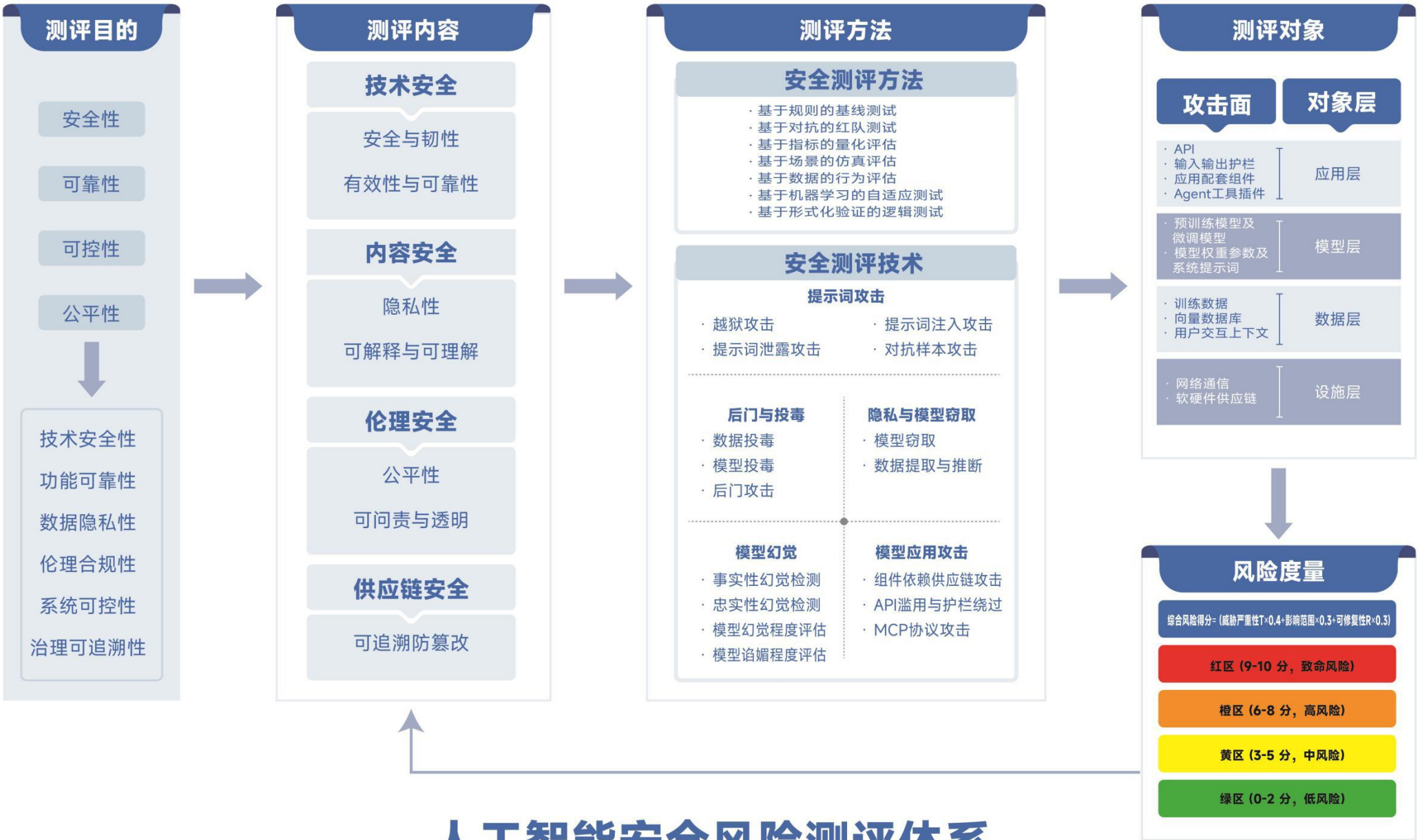
“人工智能安全测评体系”是一套基于反馈控制逻辑的全流程闭环方法，以“目标设定—内容实施—方法技术—对象覆盖—风险度量—持续优化”为链路，实现人工智能安全风险测评的系统性与动态性管理。人工智能安全风险测评体系将“目标”“对象”“内容”“方法”“度量”深度耦合，既实现测评的全维度覆盖，又通过闭环机制保障测评的动态适应性，为人工智能治理提供系统化工具。

“人工智能安全风险测评体系”核心模块有：**（1）测评目的**：锚定“安全性、可靠性、可控性、公平性”四大核心目标，为整个测评体系定义基准，明确人工智能系统需达到的安全状态，指导后续测评全流程的方向。**（2）测评内容**：从技术安全、内容安全、伦理安全、供应链安全等维度构建测评范畴，对应“有效性、可靠性、安全性、抗干扰性、可问责与透明性、可解释性、隐私增强、公平性”等具体要求，为测评提供“做什么”的内容框架。**（3）测评方法与技术**：测评方法（策略层）为测评工作提供“方法论工具包”；测评技术（执行层）聚焦提示词攻击、投毒攻击、模型窃取、规避攻击、隐私攻击等典型攻击手段的模拟与防御验证，检验系统安全韧性。**（4）测评**

对象：根据人工智能系统的分层框架，分析安全测评的攻击面，覆盖应用层、模型层、数据层以及设施层，确定人工智能系统安全风险的直接载体。攻击面聚焦查询访问权限、模型控制、标签操纵限度、训练/测试数据控制、源代码控制、资源控制等关键入口。**（5）风险度量**：基于威胁严重性、影响范围、可修复性等维度构建量化模型，将风险映射至红（致命风险）、橙（高风险）、黄（中风险）、绿（低风险）四级风险区间，精准度量测评结果与安全目标的偏差，为后续优化提供反馈信号。同时，“人工智能安全风险测评体系”还有**闭环优化**：风险度量的结果反向指导测评内容的迭代优化、针对高风险环节强化测评重点，形成“测评实施—风险度量—内容优化—再测评”的持续改进闭环。

一、测评目的：多维度安全目标

“测评目的”锚定“安全性、可靠性、可控性、公平性”四大核心目标，再依据 LLM 技术特征进行“拆解”，构建多维度安全目标体系。基本的，LLM 系统的安全测评需超越单一的“防攻击”范畴，立足技术特性、伦理要求与治理规范的交叉维度，构建覆盖全生命周期的多维度安全目标体系。这一体系既需响应国际标准，又需适配 LLM 特有的风险场景（如大模型的泛化能力带来的滥用风险），实现“技术安全、伦理合合规、治理可控”统一。基于 LLM 系统的复杂性与影响范围，核心测



人工智能安全风险测评体系

评目标细化为六大维度，各维度既独立聚焦特定安全需求，又协同支撑系统整体的可信性，其中，“**技术安全性**”“**功能可靠性**”保障“能安全工作”，“**数据隐私性**”“**伦理合规性**”确保“不造成伤害”，“**系统可控性**”“**治理可追溯性**”实现“风险可管控”。通过六大维度的协同测评，全面刻画 LLM 系统的安全状态，为人工智能安全应用提供完整的评估依据。

技术安全性是 LLM 系统的基础保障目标，核心在于抵御各类恶意攻击与技术扰动，确保系统在面临人为干预或异常环境时仍能维持核心功能的稳定性。测评内涵包括三层：一是对抗鲁棒性，即系统对对抗样本（如微调的恶意提示词、扰动的输入文本）的抵御能力，需验证模型在攻击下的错误率是否控制在可接受范围（如对抗样本成功率≤5%）；二是抗窃取能力，针对模型权重、架构等核心资产，测评其抵御提取攻击（如通过 API 查询反推参数）的防护机制；三是动态稳定性，关注模型在迭代更新（如持续微调、部署环境变化）过程中的安全退化风险（如某次微调后对抗防御能力下降）。

功能可靠性关注 LLM 系统在正常及边缘场景下输出的准确性与一致性，确保其智能能力的“可用且可信”。核心测评维度包括：基础准确性，在常规输入下的任务完成质量（如问答系统的答案正确率、翻译系统的语义一致性）；分布外鲁棒性，针对训练数据未覆盖的边缘样本（如罕见词汇、新兴概念），测评模型的泛化能力与错误控制（如“不知道时不编造”的诚

实性）；场景适配性，在特定行业场景（如医疗诊断、金融风控）中，验证模型输出与业务规则的匹配度（如医疗人工智能的诊断建议是否符合临床指南），避免因功能失效导致的业务风险。

数据隐私性聚焦 LLM 系统全链路的数据保护，防止训练数据、用户输入及模型衍生信息的未授权泄露或滥用，是合规性与用户信任的核心支撑。其核心测评方向包括：训练数据隐私，验证训练集中个人敏感信息（如身份证号、医疗记录）的脱敏效果，以及抵御“成员推理攻击”（通过模型输出判断某样本是否在训练集中）的能力；用户交互隐私，检测用户输入数据（如对话内容、查询请求）在传输、存储、处理环节的加密与访问控制措施，避免实时交互中的信息泄露；模型衍生隐私，关注模型“记忆”训练数据细节导致的风险（如大模型复述训练集中的隐私文本），需验证模型对敏感信息的“遗忘”能力与输出过滤机制。

伦理合规性旨在确保 LLM 系统的设计与输出符合社会伦理准则、法律规范及文化价值观，避免因技术滥用或偏见导致的社会风险。测评内涵需兼顾普适性与场景化：一是反歧视与公平性，验证模型对不同群体（如性别、种族、地域）的输出是否存在系统性偏见（如招聘场景中对女性候选人的评分偏低），可通过构建多样化测试集量化偏见指数；二是价值观对齐，测评模型输出与主流社会价值观（如反对暴力、尊重人权）

的一致性，尤其需适配多文化场景（如不同地区对 “言论自由”的边界定义差异）；三是法律合规性，对照区域法规（如欧盟人工智能法案禁止的“社会评分”应用、中国《生成式人工智能服务管理暂行办法》的内容安全要求），验证系统是否满足禁止性与义务性条款。

系统可控性强调对 LLM 系统行为的可预测、可干预与可终止能力，防止系统出现超出设计预期的“失控”行为。测评重点包括：目标对齐性，验证模型实际行为与预设目标（如 “安全对话”“合规生成”）的一致性，抵御 “目标偏移攻击”（如通过诱导使模型优先满足用户恶意需求）；干预有效性，测试紧急情况下的人工干预机制（如 “一键关停”“输出拦截”）的响应速度与可靠性；行为可预测性，通过大量测试样本分析模型输出的波动范围，确保无 “突发异常”（如无明显诱因的有害内容生成）。

治理可追溯性聚焦 LLM 系统全生命周期的操作可审计与责任可定位，支撑安全事件溯源与改进。核心测评内容包括：全链路日志完整性，验证模型开发（如微调参数变更）、训练（如数据来源）、部署（如版本更新）、运行（如用户交互）各环节日志的记录完整性与不可篡改性；供应链可追溯，对第三方组件（如开源框架、数据集）来源、版本及安全认证进行链式核查，确保风险可溯源；责任主体明确性，通过文档审查与流程验证，确认系统各环节的责任部门/人员（如模型开发者、

部署运维方），确保安全事件发生后可快速定位责任主体。

二、测评内容：全领域安全覆盖

基于大模型全生命周期、人工智能技术特征、全生命周期风险传导规律、落地应用的安全风险暴露面等关键支撑，构建生成式人工智能安全风险全景图。同时，因风险暴露面相似等原因，绘制“人工智能安全风险全图景”需对关键维度进行适当调整以突出测评内容：将全生命周期中的“模型训练与构建”与“模型验证与确认”合并为“模型训练优化”，将全生命周期中的“用户使用与影响”扩充为“迭代与退役”。由此，人工智能安全风险全图景既能沿用全生命周期呈现不同阶段的“异质风险”，又在基础设施安全、数据安全、模型安全、应用与智能体安全、用户与身份安全、内容安全、合规与伦理风险、管理类风险等层次呈现“同质风险”，支撑测评内容的“框定”。结合人工智能安全风险全景图，测评内容覆盖从技术安全、内容安全、伦理安全、供应链安全等维度，聚焦供应链、数据、模型自身、价值观与伦理对齐、运行态系统五大安全风险测评重点，能进一步明确核心测评维度与关键核查方向，实现风险测评从“框架”到“重点”的精准落地。从人工智能安全风险全景图提取测评重点，可用于对测评工作开展初步的规划与指导，包括供应链安全测评、数据安全测评、模型安全测评、价值观与伦理对齐测评、运行态系统安全测评。

人工智能安全风险全景图

风险对象层次	规划与设计	数据采集与处理	模型训练与优化	部署与集成	运行与监测	迭代与退役
基础设施安全	·算力架构设计缺陷 (SR) ·硬件选型安全风险 (SR) ·算力供应链风险 (SR、AC) ·云资源不可靠规划 (RE、SR)	·存储硬件漏洞 (SR、PE) ·网络传输链路风险 (SR、PE) ·内存安全威胁 (SR、PE) ·数据污染 (存储层)(RE、SR)	·训练集群权限不足 (SR) ·CPU/IGPU 安全威胁 (SR) ·开发框架风险 (SR、RE) ·资源可用性不足 (RE、SR)	·部署环境配置错误 (SR、RE) ·云 / 边缘设备固件漏洞 (SR) ·中间件安全风险 (SR) ·第三方组件不可信 (SR、AC)	·算力资源滥用 (SR、RE) ·系统日志泄露 (PE、SR) ·操作系统 / 数据库漏洞 (SR) ·网络隔离不足 (SR、PE)	·硬件数据残留 (PE、AC) ·资源回收不彻底 (SR、AC) ·供应链组件清理缺失 (SR) ·依赖组件过旧 (RE、SR)
数据安全	·数据合规性缺失 (PE、AC) ·隐私保护方案遗漏 (PE) ·跨境合规规划不足 (PE、AC) ·训练数据隐私风险规划 (PE)	·训练数据投毒 (RE、SR) ·采集阶段数据窃取 (PE、SR) ·敏感信息残留 (PE) ·知识阵投毒 (RE、SR) ·标签质量风险 (RE、FA)	·训练数据隐私泄露 (PE) ·标注错误引入偏见 (FA、RE) ·知识库信息泄露 (PE、SR) ·数据偏见与歧视 (FA)	·接口未加密 (PE、SR) ·生产 / 训练数据混流 (RE、PE) ·未授权数据访问 (SR、PE) ·数据存储传输未加密 (PE、SR)	·实时数据窃取 (PE、SR) ·用户数据过度收集 (PE、AC) ·存储层数据泄露 (PE、SR) ·个人隐私泄露 (输出)(PE)	·退役数据未脱敏 (PE、AC) ·数据销毁流程缺失 (PE、AC) ·知识版权争议 (清算)(AC) ·输出结果被恶意利用 (SR、AC)
模型安全	·架构安全设计缺陷 (SR、RE) ·安全审计标准缺失 (AC、TR) ·算法逻辑缺陷 (设计)(RE、EX) ·模型与企业风险脱节 (AC、RE)	-	·模型窃取 / 逆向工程 (SR、AC) ·训练数据偏见 (FA、RE) ·算法逻辑缺陷 (优化)(RE、EX) ·版本管理混乱 (TR、AC) ·模型投毒 (RE、SR)	·API 未授权调用 / 滥用 (SR、AC) ·模型劫持 (SR、RE) ·版本恶意替换 (SR、AC) ·部署版本不一致 (RE、TR)	·输出结果篡改 (SR、RE) ·模型漂移未监控 (RE、TR) ·时间 / 资源消耗攻击 (SR、RE) ·模型预测不可靠 (RE)	·旧模型未下架 (SR、AC) ·参数销毁不彻底 (PE、SR) ·版本管理消算缺失 (TR、AC) ·模型供应链风险清算 (AC、SR)
应用与智能体安全	·安全边界模糊 (SR、TR) ·智能体权限过度设计 (SR、AC) ·Web 应用漏洞 (设计)(SR) ·身份权限规划不足 (SR、AC)	-	-	·提示注入漏洞 (SR、RE) ·插件权限失控 (SR、AC) ·Web 应用漏洞 (部署)(SR) ·外部组件集成风险 (SR、RE) ·不安全配置 (SR、TR)	·业务流程劫持 (SR、AC) ·XSS/CSRF 漏洞 (SR) ·插件恶意行为 (SR、AC) ·智能体目标偏移 (RE、AC) ·多智能体恶意推理 (SR、RE)	·退役应用后门未关 (SR、AC) ·关联系统解绑不彻底 (SR、AC) ·智能体权限回收缺失 (SR、AC) ·依赖项安全清算 (SR、RE)
用户与身份安全	·权限体系设计缺陷 (SR、AC) ·身份认证方案薄弱 (SR、PE) ·权限过度授予 (设计)(SR、AC) ·用户身份认证风险规划 (SR、PE)	-	-	·LAM 配置错误 (SR、AC) ·多租户隔离失效 (SR、PE) ·权限过度投予 (配置)(SR、AC) ·访问控制不足 (SR)	·身份冒充 (SR、PE) ·会话劫持 (SR、PE) ·敏感信息窃取 (PE、SR) ·未授权用户操作 (SR、AC) ·智能体身份冒用 (SR、AC)	·退役账号未注销 (SR、AC) ·权限回收不及时 (SR、AC) ·用户身份信息清算缺失 (PE、AC) ·LAM 集成清算不足 (SR、AC)
内容安全	·生成边界未定义 (TR、AC) ·审核规则缺失 (TR、AC) ·法律责任风险 (规划)(AC) ·上下文不安全规划 (RE、TR)	-	-	·输出过滤机制未部署 (TR、AC) ·审核机制缺失 (部署)(TR、AC) ·过度权限工具调用风险 (SR、AC)	·色情 / 暴力 / 虚假信息 (TR、AC、RE) ·意识形态 / 伦理争议内容 (AC、FA) ·歧视性内容未拦截 (FA、AC) ·智能体利用内容欺骗 (RE、AC)	·历史内容未归档审核 (TR、AC) ·法律责任风险 (灏算)(AC) ·内器依赖项安全清算 (SR、AC)
合规与伦理风险	·未适配地区性 AI 法规 (AC、TR) ·伦理审查流程缺失 (AC、FA) ·合同合规风险 (签订)(AC) ·缺乏安全治理框架 (TR、AC) ·可解释性不足 (EX、TR)	·违反隐私法 (PE、AC) ·知情同意缺失 (PE、TR) ·数据跨境合规风险 (PE、AC) ·违反数据保护法规 (PE、AC)	·算法偏见合规评估不足 (FA、AC) ·AI 法规不合规 (训练)(AC) ·组织对 AI 风险识别不足 (TR、AC)	·未通过安全认证 (AC、TR) ·合同合规风险 (执行)(AC) ·模型合规性不足 (AC) ·不安全技术使用 (SR)	·决策透明度不足 (TR、EX) ·责任归属争议 (AC) ·伦理争议内容合规风险 (AC、FA) ·安全隐患与社会影响 (AC、TR)	·未完成合规备案 (AC、TR) ·伦理争议未闭环 (AC、FA) ·合同合规风险 (终止)(AC) ·社会影响清算不足 (AC、TR)
管理类风险	·安全责任未明确 (AC) ·风险评估流程缺失 (TR、AC) ·安全策略缺失 (TR、AC)	·数据供应链不可信 (AC、SR) ·供应商数据泄露 (AC、SR) ·第三方组件恶意植入 (SR、AC)	·训练文档不全 (TR、EX) ·操作规范缺失 (TR、AC) ·人员操作失误 (训练)(AC、RE)	·部署验收标准缺失 (TR、AC) ·供应链组件审计不足 (TR、SR) ·人员操作失误 (部署)(AC、RE)	·监控机制失效 (TR、RE) ·应急响应滞后 (AC、SR) ·安全意识不足 (AC、TR)	·退役处置流程未标准化 (TR、AC) ·知识传递断层 (TR、EX) ·责任清算流程缺失 (AC)

核心特征： PE= 隐私增强　SR= 安全与韧性　FA= 公平性　EX= 可解释性　RE= 可靠性　AC= 问责性　TR= 透明度

（一）供应链安全测评

供应链安全测评聚焦核心组件与服务在采购、研发、交付、集成等环节存在的安全风险，如，供应链攻击、组件漏洞、来源不可靠、交付链路失控等。主要测评对象包括物理硬件、底层基础设施、软件及开发框架，模型及相关核心组件，以及供应链各环节引入的第三方服务等。

1.硬件及基础设施供应链安全测评聚焦支撑大模型运行的物理硬件及底层基础环境的供应链安全，围绕硬件供应链的供应韧性、完整性与计算安全性，以及基础设施供应链的组件固有风险、供应链劫持与配置安全两大维度开展测评，识别发现硬件生产、供应、交付及底层基础设施搭建、部署、配置等供应链起点源头安全隐患，主要内容有:

硬件供应链安全测评围绕核心硬件的供应韧性、完整性与计算安全性展开。供应韧性评估关注地缘政治风险下的“断供”可能性，对国内备选供应商的产能、工艺成熟度及合规认证进行全面分析。硬件完整性与计算安全性评估，主要通过物理检测和硬件可信机制验证等手段排查硬件是否存在物理篡改、固件篡改及恶意元件植入的风险。

基础设施供应链安全测评重点关注支撑大模型运行的底层环境组件的固有风险、供应链劫持及配置安全。组件固有风险评估需深入到固件层面，检测 BIOS/UEFI 等固件是否存在已知

漏洞。供应链劫持风险评估主要关注容器镜像仓库投毒等常见威胁，以及基础设施即代码模板的安全性，防止恶意配置通过自动化部署流程扩散。配置安全风险评估主要关注因供应商在开发或交付阶段预设的不当配置引入的风险。

2.软件及开发框架供应链安全测评针对大模型开发与运行所依赖的基础软件、开发框架、组件及工具链，从核心开源组件风险、供应链流转过程完整性、合规与安全认证三个维度评估恶意组件植入、漏洞传递、合规缺失等风险，主要内容有:

核心开源组件风险测评从供应链源头识别安全隐患，针对开源社区主导研发或维护的人工智能框架、开发库及中间件，评估其安全漏洞修复响应效率、社区维护规范性、安全测试覆盖深度，自主可控程度以及国产化适配能力。

供应链全流程完整性测评重点关注软件及开发框架从代码提交、编译、打包到部署全流程的防篡改与可追溯能力。在研发环节，核查代码版本控制系统的访问权限管控与签名提交机制，评估代码评审流程规范性及高危代码修改审批复核机制有效性；在构建部署环节，核验持续集成/交付流水线各节点安全管控措施，编译、打包、镜像构建等环节产物的哈希值或数字签名生成情况，以及 SLSA 证据链的完整性与可审计性，评估从源码到部署包的每一步转换是否可追溯、防篡改。

合规与安全认证情况核验主要针对第三方商业软件，核查授权文件合法性与有效性，是否通过国家权威安全认证，并核

查供应商提供软件物料清单的完整性、时效性与规范性，评估供应链的透明度与风险响应的敏捷性。

3.模型供应链安全测评主要针对大模型本体以及其训练、微调、部署、迭代全生命周期各环节的供应链安全开展测评，主要内容有:

模型来源可信性测评聚焦模型源头，核查研发主体资质合规性、技术自主化程度。针对采用外部供应模型的场景，重点核验供应商提供信息的完整性与真实性，具体涵盖两方面内容:一是模型核心属性信息，包括性能指标、训练数据来源及合规性证明、功能局限性、潜在偏见与伦理风险说明等；二是模型安全相关信息，包括功能模块披露完整性、数据处理机制透明度、隐私保护措施说明等，通过上述核验评估模型来源可追溯性与属性可核查性。

模型血缘与物料追溯测评聚焦模型供应链的血缘透明度与风险继承可追溯性。核心测试内容包括：一是供应商是否完整列明预训练模型的构成要素；二是模型从基础版本到最终交付版本的迭代链路是否完整，识别评估模型可能继承的上游风险。

4.第三方服务供应链安全测评针对大模型运行依赖的第三方服务，围绕“第三方服务依赖风险”，从服务商准入、服务接口、故障应急、持续管控四个方面评估第三方供应链安全风险。

（二）数据安全测评

数据安全测评聚焦数据全生命周期安全风险，针对数据本身及相关处理活动，从训练数据来源及内容安全、数据存储及传输安全、数据使用安全、数据销毁安全四个维度评估数据机密性、完整性、合规性等方面的综合防护能力。

1.训练数据来源及内容安全测评针对大模型训练数据来源及内容的安全合规性，从训练数据来源及流转安全、训练数据本身的内容安全两方面开展测评，主要内容有:

训练数据来源及流转安全测评聚焦于数据从外部获取到进入训练环节的源头安全，重点关注训练数据来源合规和数据进入训练之前的流转安全两个方面。训练数据来源合规测评重点关注训练数据是否来自正规授权渠道或公开合法数据集；训练数据流转安全测评重点关注数据流转全链条安全管控机制的有效性。

训练数据内容安全测评聚焦于训练数据内容本身的安全性与合规性，识别训练数据中潜藏的敏感信息与恶意内容，主要包括：一是敏感信息残留评估，二是数据污染风险评估，三是训练数据合规冲突评估。

2.数据存储及传输安全测评聚焦数据存储与传输环节数据安全防护措施的有效性，围绕数据完整性和保密性开展安全评估，主要内容有:

数据存储安全测评聚焦数据静态存储阶段数据安全防护措

施的有效性。一是核验本地服务器、云存储节点、分布式数据库等存储介质的静态加密机制，二是评估存储系统的访问控制策略。

数据传输安全测评聚焦数据动态流转全链路数据安全防护措施的有效性。一是关注传输协议的安全性，二是评估跨域传输时数据脱敏处理效果，以及网关设备、API 接口、服务器端口、终端接入设备等各类传输接入节点的身份认证强度。

3.数据使用安全测评聚焦人工操作、系统处理、模型运行对数据的全场景使用行为，主要包括数据操作合规测评和敏感数据泄露测评，主要内容有：**数据操作合规测评**针对数据使用过程的合规性开展评估，关注数据操作行为是否全程留痕、责任可究，核验数据访问日志的完整性与可审计性，评估数据使用的合规可控程度；**敏感数据泄露安全测评**聚焦评估模型在训练和推理过程中对训练数据的保护能力，核心测试内容围绕训练数据提取及成员推理攻击下的脆弱性展开。

4.数据销毁安全测评聚焦数据生命周期末端数据使用完毕后物理销毁与逻辑销毁的合规性、彻底性及可追溯性，评估废弃数据被非法恢复、窃取引发的安全风险，形成数据全生命周期安全测评闭环。

（三）模型安全测评

模型安全测评聚焦于模型架构、参数、推理机制等内在属

性所固有的安全风险，从模型对抗鲁棒性、输出可靠性、模型完整性与后门风险三个维度测评大模型自身的潜在安全风险。

1.对抗鲁棒性测评重点关注大模型在面对非标准输入、恶意攻击等复杂条件时，能否保持输出稳定性以规避风险。该维度测评重点考察大模型对提示词攻击、对抗性样本攻击等典型威胁的防御能力。对抗性鲁棒性测评还需扩展到其他模态。针对多模态模型，还需测试其在图文结合输入中面对误导性图像时的鲁棒性，防止跨模态攻击导致误判。例如，评估模型能否抵御隐藏在图像像素中的对抗性扰动，或者能否识别并拒绝执行嵌入在图片或音频中的不可见的恶意指令。

2.输出可靠性安全测评关注大模型生成内容是否真实、可信、可解释、可追溯，聚焦大模型幻觉抑制能力与内容可解释性，防范因虚假或不可控输出引发的信息误导与决策风险，主要内容有：**幻觉与真实性测评**针对大模型生成“与事实不符、无依据的虚假信息”的幻觉问题展开，判断其生成内容是否符合事实；**可解释性与溯源能力测评**主要关注大模型生成内容的推导逻辑透明度与依据可追溯性。

3.模型完整性与后门安全测评聚焦于评估模型本身的完整性是否遭到破坏，以及是否被植入了隐藏的、恶意的后门，主要内容有：**模型完整性测评**重点评估模型发布方的身份认证机制、分发渠道安全性及文件完整性保护措施。测评验证模型权重文件是否采用数字签名、是否提供可校验的哈希值，确保传

输过程未被篡改。对于开源模型，检查代码仓库是否启用代码签名、持续集成/交付流水线是否集成安全扫描，防止恶意代码注入。**后门安全测评**需要审查训练流程的公开信息，包括数据预处理逻辑、清洗策略与来源清单，并通过特定技术手段重点检测是否存在“数据毒化”行为。同时，评估模型是否具备训练数据溯源与可疑数据隔离能力。

（四）价值观与伦理对齐测评

价值观与伦理对齐测评聚焦生成式人工智能的价值观输出、伦理遵循是否符合法律法规、社会规范以及公序良俗等，主要从内容合规性、偏见与公平性、伦理与道德遵从、社会与文化影响、意识形态安全五个方面开展测评。

1.内容合规性安全测评聚焦模型输出内容的“无害性”，评估模型对违规内容的识别、过滤及抵制能力。

2.偏见与公平性测评聚焦模型输出内容的“公平、平等、反歧视”特性，核心测试内容主要包括：一是群体刻板印象与歧视规避测试，二是决策机会公平性保障测试，三是文化多样性与包容性尊重测试。

3.伦理与道德遵从测评重点围绕大模型对“尊重隐私、诚实守信、人文关怀”等伦理准则的价值倾向，主要关注：一是隐私边界识别与隐私保护，二是诚信与知识产权保护，三是人文关怀与向善引导。

4.社会与文化影响测评聚焦生成式人工智能对社会环境、文化传承发展的潜在外部影响，核心测试内容主要包括：一是社会信任维护测试，二是文化安全影响测试。

5.意识形态安全测评聚焦生成式人工智能自身的价值导向，在输出内容合规的基础上，以更严格明确的标准划定其意识形态安全边界，核心测试内容主要包括：一是政治安全层面，二是历史虚无主义抵制层面，三是主流价值观遵循层面。

（五）运行态系统安全测评

运行态系统安全测评聚焦大模型部署应用后，集成应用系统依托的计算平台、存储、网络及安全护栏等所采取的安全措施的整体防护效果。测评核心围绕“基础设施—网络环境—应用载体—智能体”全栈安全体系，关注各层面安全措施的有效性。

1.基础设施安全测评聚焦大模型集成应用系统所依托的计算平台与存储设施，重点核查硬件级与系统级安全防护能力，主要内容有：**计算平台安全测评**主要关注服务器、容器集群等核心计算设备在访问控制、安全加固、资源调度安全等方面的安全防护能力。**存储设施安全测评**聚焦数据存储的可靠性保障，主要关注数据访问审计和容灾备份。在数据访问审计方面，需核查存储介质的访问日志记录的完备性和留存时长，并验证日志的不可篡改性，评估在发生数据安全事件时能否实现完整追

溯。在容灾备份方面，重点关注备份策略、数据完整性、恢复有效性三个方面。

2.网络安全测评聚焦大模型集成应用系统所处网络环境的边界防护强度、通信安全可靠性，以及网络行为审计等方面的安全防护能力，重点验证网络层面抵御外部攻击渗透的能力，主要内容有：**边界防护安全测评**核心围绕网络入口与区域隔离的安全管控能力，主要关注网络隔离、服务端口管控、防护设备效能三个方面。**通信安全测评**聚焦大模型数据传输全链路的加密防护能力。主要关注传输加密及密钥管理等安全机制的落地效果。**网络行为审计测评**聚焦网络访问行为的可追溯性与异常监测能力，核心关注日志管理、日志分析能力两大维度。**应用安全测评**聚焦大模型集成应用实际运行中业务逻辑、服务接口、代码实现等关键环节的安全风险，重点关注业务逻辑安全测评、服务接口安全测评及代码安全审计三大维度。

3.智能体安全测评聚焦人工智能智能体在“思考—行动”循环中所引入的独特风险，关注工具调用与执行安全、任务规划与分解可靠性及外部信息源的完整性与真实性三大维度，主要内容有：

工具调用与执行安全测评重点评估工具调用的安全边界，主要包括：一是核查智能体可调用的工具列表是否遵循“最小权限”原则，是否存在不必要的、高风险的工具授权；二是评估对工具输入参数的净化与验证机制，防止恶意构造的参数在

执行时导致命令注入或非预期行为。三是检验工具执行环境的沙箱化程度能否确保即使工具被恶意利用，其影响也被严格限制在隔离环境中，无法危害宿主系统。

任务规划与分解的可靠性测评重点关注其任务规划逻辑的可靠性与安全性，主要包括：一是任务分解逻辑方面，通过对抗性测试，测评智能体在面对模糊、矛盾或带有恶意意图的指令时，其任务分解过程是否会出现逻辑漏洞，如产生无限循环、执行破坏性操作序列或泄露敏感信息等。二是长期任务状态管理方面，关注智能体在长期任务执行过程中的状态管理，其状态是否会被篡改，以及被篡改后是否会导致任务偏离预期轨道，评估其状态管理机制能否保障任务执行的准确性。

外部信息源的完整性与真实性测评重点关注智能体对外部信息源的验证能力，如智能体是否会盲目信任并执行从不可信网站或文档中获取的指令，或者是否会基于虚假信息做出错误的决策和行动，是否存在信息来源验证、交叉比对等机制，以增强智能体对外部信息的“免疫力”。

三、测评方法：多元化技术路径

LLM 系统的安全风险具有复杂性、动态性与场景依赖性，单一测评方法难以覆盖其全维度安全需求。基于“技术特性—风险类型—应用场景”的匹配逻辑，需构建多元化测评技术路径，融合规则化、对抗性、量化分析、场景仿真、智能自适应

及形式化验证等方法，形成“静态筛查—动态攻击—量化评估—场景验证”的闭环体系。不同方法既各有侧重（如规则方法聚焦合规底线，对抗方法聚焦攻击面挖掘），又能协同互补（如量化方法为对抗测试提供结果锚点，场景仿真为规则库提供更新依据），最终实现对 LLM 系统安全状态的全面、精准刻画。

多元化测评技术路径的核心价值在于“优势互补、场景适配”，主要涉及：**基于规则的方法**筑牢合规底线，**基于对抗的方法**挖掘漏洞，**基于指标的方法**实现量化对比，**基于场景的方法**贴近业务风险，**基于监测的方法**偏向长周期，**基于机器学习的方法**提升测评效率，**基于形式化验证的方法**保障核心安全。在实际测评中，需根据 LLM 的应用场景（如通用 vs 行业）、安全目标（如合规 vs 鲁棒性）及资源约束（如时间、成本），灵活选择单一方法或组合策略（如“规则筛查+对抗测试+场景仿真”的组合适用于大多数行业应用），最终实现全面、精准、高效的安全评估。

（一）基于规则的基线测试

基于规则的基线测试是 LLM 安全测评的基础方法，核心逻辑是将政策标准、伦理准则、业务规范转化为可执行的刚性规则，通过“符合性校验”快速判断系统是否满足安全底线要求。技术原理上，需先构建多维度规则库：政策合规规则（如欧盟人工智能法案中“禁止生成社会评分内容”的条款、中国《生

成式人工智能服务管理暂行办法》中“训练数据合规性要求”）、伦理准则规则（如 UNESCO AI 伦理中“反对歧视”的具体判定标准、行业伦理规范如医疗人工智能的“患者利益优先”原则）、技术安全规则（如“禁止响应恶意指令”的关键词库、“用户数据加密传输”的协议要求）。测评过程中，通过自动化工具将 LLM 系统的配置参数、输出内容、操作日志与规则库进行匹配，输出“符合/不符合”的明确结论。

该方法的适用场景集中在合规性快速筛查与固定风险点检测，如，对新上线的生成式人工智能应用进行内容安全初审（通过关键词匹配检测是否生成暴力、色情内容），对模型训练数据进行合规性校验（核查是否包含未授权的个人信息）。基于规则的基线测试优势在于实施门槛低、结果可解释性强（不符合项可直接定位至具体规则），适合作为测评的“第一道防线”。但局限性也较为明显：规则库依赖人工预设，难以覆盖复杂语义场景（如隐喻式有害内容）或新兴风险（如新型提示词攻击），需结合其他方法弥补灵活性不足的问题。

（二）基于对抗的红队测试

基于对抗的红队测试是主动挖掘 LLM 系统脆弱性的重要方法，核心逻辑是模拟攻击者视角，通过系统化的对抗技术触发系统安全漏洞，验证其防御机制的有效性。技术原理上，需构建“攻击策略库”与“红队操作流程”：攻击策略库涵盖 LLM

特有的攻击手段（如提示词注入、角色越权、多模态对抗样本生成等），红队操作流程则包括目标分析（明确测评对象的攻击面，如 API 接口、护栏机制）、攻击实施（按严重程度分级测试，从简单攻击到复杂组合攻击）、结果记录（记录漏洞触发条件、影响范围、复现路径）。例如，针对 LLM 的护栏机制，红队可通过“逐步诱导式提示词”（先建立信任，再隐蔽植入恶意指令）测试其被绕过的可能性；针对模型鲁棒性，可通过 GAN 生成语义相近但带有扰动的文本，测试模型输出错误率的变化。

该方法的适用场景集中在模型鲁棒性测评、护栏机制有效性验证及未知漏洞挖掘，如，评估大模型对“越狱攻击”（Jailbreak Attacks）的防御能力，验证 RAG 系统对知识库污染攻击的抵御效果。基于对抗的红队测试优势在于贴近真实场景，能发现规则测试难以覆盖的“隐性漏洞”（如逻辑绕过、条件触发漏洞），为系统加固提供直接依据。但实施需依赖专业红队人员（具备人工智能安全与攻击技术复合能力），且测试结果受攻击策略完备性影响较大，需结合最新攻击研究动态持续更新策略库。

（三）基于指标的量化评估

基于指标的量化评估是将 LLM 安全状态“可感知、可比较”的关键方法，其核心逻辑是通过客观、可计算的指标体系，将定性的安全描述转化为定量数据，实现不同系统、不同阶段的

安全水平对比。技术原理上，需构建“多层级指标库”：基础指标（如对抗样本成功率 ASR、有害内容拦截率 HCR、偏见指数 DI 等）用于刻画单一维度的安全状态；综合指标（如风险等级得分）通过加权算法（如层次分析法 AHP）整合基础指标，反映系统整体安全水平。测评过程中，先通过规则测试、对抗测试等方法获取原始数据，再代入指标计算公式输出量化结果，如，某 LLM 的对抗样本成功率为 8%（绿区），偏见指数为 1.2（黄区），综合风险等级为 4.5 分（中风险）。

该方法的适用场景集中在安全状态的横向对比（如不同厂商 LLM 的鲁棒性比较）、纵向追踪（如同一模型迭代前后的安全改进幅度）及风险优先级排序，如，企业在选型 LLM 时，通过“模型窃取防御成功率”“用户数据泄露风险值”等指标量化评估候选模型；监管机构通过“合规达标率”指标对辖区内人工智能应用进行分级监管。基于指标的量化评估优势在于结果客观、可操作性强，为决策提供数据支撑。但需注意指标的科学性：一是指标定义需贴合 LLM 特性（如“分布外鲁棒性”指标需区分文本、多模态场景）；二是权重设置需结合应用场景（如金融领域“数据隐私性”指标权重应高于通用场景）。

（四）基于场景的仿真评估

基于场景的仿真评估是验证 LLM 系统在实际业务中安全表现的重要方法，核心逻辑是构建贴近真实应用的场景库，通

过“沉浸式测试”发现场景化特有风险。技术原理上，需分三步实施：**场景建模**，梳理目标 LLM 的典型应用场景，如“金融风控人工智能的贷款审批”“医疗人工智能的辅助诊断”“教育人工智能的个性化辅导”，明确场景中的角色（用户、系统、第三方）、流程（输入-处理-输出）及安全诉求（如金融场景需防欺诈、医疗场景需保隐私）；**测试用例生成**，基于场景特征设计高覆盖度的测试用例，如金融场景的“异常收入证明材料审核”“敏感人群贷款申请处理”；**仿真执行**，在模拟环境中复现业务流程，记录系统在场景化输入下的安全表现，如是否泄露申请人隐私、是否对低收入群体存在偏见）。

该方法的适用场景集中在行业专属人工智能应用的安全测评，如，测评医疗 LLM 在“罕见病诊断”场景中是否泄露患者病历信息、是否因训练数据不足导致误诊风险；测评司法 LLM 在“量刑建议”场景中是否对特定罪名存在量刑偏差。基于场景的仿真评估优势在于突破通用测评的“抽象性”，直接关联业务风险（如误诊可能导致的医疗纠纷），为行业客户提供“可落地”的安全评估结论。但场景库构建成本较高，需联合行业专家（如医生、法官）参与设计，且场景需随业务迭代动态更新（如金融场景新增“数字人民币交易风控”）。

（五）基于数据的行为评估

基于数据的行为评估是验证 LLM 系统在实际业务中的行

为符合人类的价值观、意图与社会规范，避免产生偏见、歧视、有害甚至危险的输出，核心逻辑是建立一套科学、系统、有效的大模型行为评估体系，通过 LLM 系统行为数据分析特有风险。此外，相关研究指出大模型能“觉察”测试评估环境，选择“装傻充愣”，为达成目的故意答错问题、暗中修改数据或者操作记录掩盖违规行为。该方法监测大模型行为数据，完成以下工作：安全性评估，防止模型产生直接的人身或社会危害，如，生成暴力、仇恨言论、违法信息、虚假信息或提供危险建议；可靠性评估，检验大模型输出的准确性、一致性、逻辑性，避免“一本正经地胡说八道”，确保在专业、严肃场景下的可信度；公平性与包容性评估，检测并消减模型因训练数据而产生的对特定性别、种族、地域、文化群体的系统性偏见与歧视，促进技术普惠；价值观对齐评估，确保模型的行为符合社会价值观或组织所倡导的积极、健康、有益的导向，而非价值中立或价值虚无。

该方法的适用场景集中在国计民生相关领域，目的是将人类社会价值观、道德与智慧，转化为人工智能可理解、可执行的代码与约束，确保人工智能始终服务于人类福祉，如，金融（反欺诈监测等）、医疗（伦理安全等）、教育（内容适龄性评估等）、政务（抗压能力测试等）。基于数据的行为评估需要技术、伦理与治理的深度融合，面临人类价值观和社会规范复杂性、价值观跨文化冲突、标准化与一致性等方面的挑战。

（六）基于机器学习的自适应测试

基于机器学习的自适应测评是应对 LLM 规模扩大与风险动态演化的智能化方法，核心逻辑是利用人工智能技术自动生成测试用例、优化测试策略，实现对大规模、复杂系统的高效测评。技术原理上，依托两类核心技术：一是测试用例自动生成，通过生成式模型（如 GAN、LLM 自身）生成多样化测试样本（如对抗性提示词、边缘输入文本），覆盖传统人工难以穷尽的测试空间；二是测试策略自适应优化，通过强化学习（RL）训练“测评智能体”，根据系统反馈（如漏洞触发结果）动态调整测试方向（如从“基础提示词攻击”转向“组合攻击”），提升漏洞挖掘效率，如，针对千亿参数 LLM，可通过强化学习智能体自主探索高成功率的越狱攻击路径，比人工测试效率提升 10 倍以上。

该方法的适用场景集中在大规模 LLM 的高效测评、未知威胁的探索性测试及持续监测，如，对预训练大模型进行上线前的全面漏洞扫描；对已部署 LLM 进行周期性安全巡检，捕捉因模型迭代产生的新风险。基于机器学习的自适应测评优势在于突破人工测试的效率瓶颈与认知局限，能快速适应 LLM 技术的快速迭代。但局限性在于依赖高质量的训练数据（如历史漏洞案例），且生成的测试用例可能存在“无效样本”（如重复或无意义的输入），需结合人工校验提升精准性。

（七）基于形式化验证的逻辑测试

基于形式化验证的逻辑测试是保障高可靠性 LLM 系统安全的严谨方法，核心逻辑是通过数学建模与逻辑推理，证明 LLM 系统的核心安全属性（如“绝不生成有害内容”）在所有可能输入下均成立。技术原理上，需将 LLM 的安全规则转化为数学公式（如用谓词逻辑定义“有害内容”的判定条件），通过符号执行（将模型输入抽象为符号变量，追踪输出与符号变量的逻辑关系）、定理证明（利用数学公理验证安全属性的必然性）等技术，验证系统是否存在逻辑矛盾（如“在某类输入下，系统同时满足‘合规’与‘违规’条件”）。例如，对自动驾驶 LLM 的决策逻辑，可通过形式化验证证明“在任何情况下，模型不会输出‘优先撞向行人’的指令”。

该方法的适用场景集中在高风险领域 LLM 的核心模块测评，如，自动驾驶、工业控制、航空航天等领域，其系统失效可能导致人身伤亡。基于形式化验证的逻辑测试优势在于结论具有“数学严谨性”，能从理论上排除系统性逻辑漏洞，比传统测试方法更彻底。但局限性明显：一是对复杂 LLM（如千亿参数模型）的形式化建模难度极大，目前仅适用于核心子模块（如护栏规则引擎）；二是验证过程计算成本高，难以大规模推广，需与其他方法配合使用。

四、测评对象：系统全栈分层

大语言模型系统作为复杂的智能技术综合体，其安全风险并非孤立存在于某一环节，而是贯穿于从底层基础设施到上层业务应用的全链条，拆解为设施层、数据层、模型层、应用层四个核心层级。

设施层是 LLM 系统运行的物理与环境基础，主要包含支撑模型训练、推理及部署的全部基础设施，具体涵盖物理硬件（如 GPU/TPU 服务器、存储设备）、云环境（如公有云/私有云资源池、容器化平台）、网络设施（如交换机、防火墙），以及构成系统基础的软硬件组件（如芯片固件、操作系统、开源框架依赖）。从安全测评视角看，设施层的核心关注方向包括三方面：一是物理安全，即硬件设备的物理访问控制（如机房门禁、设备防盗）与环境稳定性（如温度、湿度监测）；二是网络安全，涉及数据传输加密（如模型参数传输的 TLS 协议合规性）、网络边界防护（如 DDoS 攻击抵御能力）及内部网络隔离（如训练环境与推理环境的逻辑隔离）；三是供应链安全，重点检测硬件供应链（如芯片后门、固件漏洞）与软件供应链（如开源框架的漏洞、第三方组件的恶意篡改）中的潜在风险，确保基础设施本身的可信性。

数据层是 LLM 系统中数据的采集、存储、处理与流转的核心枢纽，是连接设施层与模型层、应用层的关键环节，具体涵盖训练数据集（如原始语料库、标注数据）、用户交互数据（如对话记录、输入查询）、模型训练过程中的中间数据（如特征

向量、损失函数值）以及模型输出结果（如生成文本、代码片段）等全生命周期数据。数据层安全直接关系到用户隐私保护、商业机密维护及系统合规性，是 LLM 应用安全风险的集中爆发点。从安全测评视角看，数据层的核心关注方向包括三方面：一是数据隐私保护，即确保用户输入数据、对话记录等敏感信息在存储与传输过程中实施有效脱敏与加密处理，防止因数据泄露导致的个人隐私侵犯或商业机密外泄；二是数据完整性保障，验证数据在传输、存储及处理过程中是否遭受未授权篡改，特别是针对训练数据投毒、模型后门植入等攻击手段，确保数据的可信度；三是数据生命周期管理，评估从数据采集、处理、存储到销毁的全链条安全管理机制，包括数据留存策略合规性、敏感数据自动识别与处理能力、以及数据访问权限的精细化控制（如基于角色的访问控制 RBAC）。

模型层是 LLM 系统的核心与灵魂，聚焦于承载智能能力的算法模型及其核心要素，具体包括预训练模型（如基础大模型的架构与权重）、微调模型（如针对特定场景优化的定制化模型）、模型参数（如注意力机制参数、激活函数阈值）及模型训练过程中形成的中间成果（如检查点文件、训练日志）。作为 LLM 智能性的直接载体，模型层的测评重点围绕其技术安全性与可靠性展开：其一，鲁棒性，即模型在面对异常输入（如噪声文本、歧义表达）或恶意扰动（如对抗性样本）时保持输出稳定性的能力；其二，可解释性，指模型决策逻辑的透明度，

能否清晰追溯输出结果的生成依据（如关键特征权重、推理路径），尤其在高风险场景（如金融决策、医疗诊断）中需满足可解释性要求；其三，对抗攻击防御能力，针对模型窃取（如通过 API 查询反推权重）、模型投毒（如训练数据污染导致的后门触发）等特有攻击手段，测评模型的先天防御机制与后天加固效果。

应用层是 LLM 系统与用户及业务场景直接交互的接口，涵盖基于模型开发的各类终端应用（如智能聊天机器人、内容生成工具、代码辅助系统）及支撑应用运行的配套组件（如检索增强生成系统 RAG、输入输出护栏机制、用户交互接口 API）。由于直接面向实际业务，应用层的安全风险与业务影响关联最紧密，其测评重点集中在功能安全与数据交互安全两大维度：功能安全聚焦应用能否在业务场景中稳定履行安全职责，如护栏机制对有害请求（暴力、歧视、虚假信息）的拦截率、RAG 系统对知识库污染内容的过滤能力；数据交互安全则关注用户与系统交互过程中的数据保护，包括用户输入信息的脱敏处理（如个人敏感信息加密）、会话数据的存储安全（如访问权限控制、数据留存合规性），以及应用与模型层、设施层之间数据传输的完整性与保密性（如防止 API 调用中的数据泄露），数据层作为数据流转的核心枢纽，确保交互数据在传输、处理与存储环节的端到端安全。

整体上，LLM 系统的全栈分层测评对象既各有侧重，又相

互关联：设施层为数据层、模型层与应用层提供安全底座，数据层作为连接枢纽确保数据在各层间安全流转，模型层的安全特性直接决定应用层的风险上限，应用层的交互安全则是用户感知系统安全性的直接窗口。

五、测评度量：风险等级划分

LLM 系统的安全还测评需通过可量化、可比较的指标体系，将多维度的安全状态转化为清晰的风险等级，为安全决策提供客观依据。基于“基础指标量化—综合维度加权—风险等级映射”逻辑，构建“多层级测评指标体系”：基础指标聚焦单一安全维度的可直接测量值，综合指标通过加权算法整合基础指标形成整体风险评分，最终映射为风险等级（如，绿区：低风险—黄区：中风险—橙区：高风险—红区：致命风险）。测评指标体系通过“基础指标量化细节、综合等级把握全局、补充指标完善治理”的三层设计，实现对 LLM 系统安全状态的全面刻画。

（一）基础指标：单一维度的量化测量

基础指标是风险评估的“原子单元”，针对设施层、数据层、模型层、应用层的核心安全维度，定义可直接计算、可重复验证的量化指标，为综合风险评估提供数据支撑。在实际中，还需明确各指标的计算方式、参考阈值等关键量化参数。

设施层安全指标涉及：**供应链漏洞密度**衡量硬件/软件供应

链中漏洞的分布强度，反映供应链源头风险；**网络攻击抵御率**评估网络边界防护对典型攻击（如 DDoS、端口扫描）的拦截能力。

数据层安全指标涉及：**抗事实性幻觉率**衡量模型在具有逻辑缺陷、事实偏差等潜在错误引导的幻觉程度评估数据集下输出正确答案的概率，反映模型的抗幻觉能力；**成员推理准确率**评估攻击者通过模型输出推断某样本是否在训练集中的成功率，反映数据隐私保护能力。

模型层安全指标涉及：**对抗样本成功率**衡量模型在对抗样本攻击下输出错误结果的概率，反映模型鲁棒性；**偏见指数**量化模型对不同群体（如性别、种族）的输出差异，反映伦理公平性。

应用层安全指标涉及：**护栏拦截率**评估应用层护栏机制对有害请求（如暴力、违法指令）的拦截能力；**API 滥用防御率**衡量 API 接口对未授权访问、批量恶意调用等攻击的防御效果。

（二）综合风险等级：多维度量化评分

综合风险等级是对 LLM 系统整体安全状态的总结性评价，通过加权整合基础指标，结合“威胁严重性、影响范围、可修复性”三维度，形成量化评分，并映射为四级风险等级，具体如下：

三维度加权模型涉及：**威胁严重性**评估漏洞被利用的可能

性及攻击技术门槛（如自动化工具可利用的漏洞严重性高于需人工定制的攻击）。**影响范围**评估漏洞触发后影响的对象与程度（如仅影响单个用户 vs 影响社会公众）；**可修复性**评估漏洞修复的难度与时间成本（如配置调整 vs 模型重构）。

四级风险等级定义涉及：**绿区（低风险）**：漏洞数量少，威胁严重性低，影响范围局限（如个别场景的轻微偏见），可快速修复（如调整 API 限流参数）。**黄区（中风险）**：存在潜在漏洞，需特定条件触发（如特定提示词可绕过部分护栏），影响范围限于业务内部（如某功能模块响应延迟），修复需针对性措施（如更新对抗样本检测算法）。**橙区（高风险）**：漏洞可被常规攻击手段利用，影响核心功能（如模型可被窃取、护栏拦截率 85%），可能导致业务中断或合规风险（如违反《生成式人工智能服务管理暂行办法》），修复需投入较多资源（如模型局部重构）。**红区（致命风险）**：存在致命漏洞，可被轻易利用（如公开工具可触发），影响范围涉及社会公众或触犯法律（如生成违法内容、大规模数据泄露），修复难度极大（如需重新训练模型）。

在“人工智能安全测评体系”中，“风险度量”列举了风险度量的计算公式，即：综合风险得分=（威胁严重性 T × 0.4+影响范围 × 0.3+可修复性 R × 0.3）”，用于支撑红（致命风险）、橙（高风险）、黄（中风险）、绿（低风险）四级风险分级。这是一个典型的线性加权模型，仅用于说明基本原理。实际中，

该模型在安全领域有效性不足，如，一个漏洞如果“威胁严重性”是致命的（T=10），即便“可修复性”很容易（R=1），其综合风险也不应被拉低。正因为如此，ISO/IEC 42005 等使用风险矩阵或阈值触发机制，而非单纯的线性加权。具体到实际，人工智能安全风险度量可以采用非线性惩罚项或最大值主导的评估模型，若任一维度的风险超过阈值（如致命风险），则整体风险直接定级为最高，不再进行加权平均，如，综合风险得分 = Max(威胁严重性 T, 影响范围 I, 可修复性 R) × alpha + Mean(威胁严重性 T, 影响范围 I, 可修复性 R) × (1-alpha)，其中 alpha 为惩罚系数（如 0.7），以突出最大短板。或者，直接采用矩阵法：将风险定级为威胁严重性 T、影响范围 I、可修复性 R 中最高等级所对应的级别，除非三者均为低。

（三）补充指标：治理与可追溯性评估

除技术安全指标外，治理与可追溯性是 LLM 系统长期安全的重要保障，需通过补充指标评估，如，**审计覆盖率**是全生命周期中可审计操作（如训练数据变更、模型版本更新、用户交互）的占比，反映风险溯源能力；**责任定位时间**指在安全事件发生后，明确责任主体（如开发者、运维方）所需的时间，反映责任机制有效性。

测评指标的价值在于指导实践，可设想的应用场景，如，**企业自查**：通过基础指标量化安全短板（如“护栏拦截率 92%”

提示需优化内容审查规则）；**第三方认证**：以综合风险等级作为安全认证依据（如红区系统不可通过认证）；**监管评估**：通过行业均值对比（如金融行业对抗样本成功率均值 5%）识别高风险企业等。在上述应用场景中，均存在动态调整的情况，如，随着 LLM 技术演进(如多模态模型普及)，需新增适配指标(如跨模态对抗样本成功率）；政策更新（如欧盟人工智能法案修订），需调整合规相关阈值（如偏见指数参考标准）等。因此，需在给定时间周期内结合技术前沿与政策变化更新指标体系，确保其持续适配 LLM 安全测评需求。

第四章　人工智能安全风险测评关键技术

人工智能系统的本质是信息系统，继承并拓展传统信息系统的核心功能，即对数据的采集、传递、处理和反馈。因此，通用的网络安全风险评估的方法和技术，仍适用于人工智能系统安全测评。同时，人工智能系统通过复杂算法和庞大参数实现从数据到智能的质变，又需要开发专门的风险评估方法与技术，完成安全风险测评工作。“测评方法与技术”在“人工智能安全测评体系”中处于核心位置，其中，“**测评方法**”聚焦人工智能安全测评策略，是界定人工智能安全测评的“方法论工具包”，而“**测评技术**”关注人工智能安全测评执行，除通用的网络安全风险评估技术外，还包括红队技术体系。整体上，面对人工智能风险特征与影响，需同时推进“测评方法”和“测评技术”研究，才能全面覆盖“测评内容”，有效测试“测评对象”，支撑“人工智能安全风险测评体系”。本章着重论述适用人工智能安全风险测评的**红队测试技术**。红队测试技术是LLM安全测评技术重要组成，通过模拟攻击者的视角、手段与策略，系统性主动探测LLM系统的脆弱性。实践中，该体系采用对抗性思维，兼顾通用性与LLM特异性，整合LLM特有的测试模式（如提示词测试、多模态协同测试），覆盖输入、训

练、模型、输出、部署等层次。在论述上，每个层次的测试技术既包含基础性红队测试技术，又关注前沿变异手段，为红队测试提供可操作、可复现、可量化的指引。

一、输入层测试：针对用户输入与外部数据

输入层是 LLM 与外界交互的第一道接口，也是最易被攻击的环节之一。攻击者通过篡改或构造输入的文本或图像数据，利用大模型的能力与安全目标的冲突或者大模型安全训练泛化能力不足的缺陷，诱导模型违反安全规则，输出错误、有害内容或泄露敏感信息。输入层红队测试技术聚焦“输入操纵”，目标是绕过前端防御（如输入过滤）或直接干扰模型决策，主要有越狱测试、提示词注入测试、提示词泄露测试、对抗样本测试等几类。

（一）越狱测试

越狱测试（Jailbreak Attacks），主要有针对大语言模型的越狱测试技术-LLM 和针对视觉语言模型的越狱测试技术-VLM 两类。

1. 越狱测试技术-LLM

越狱测试技术-LLM 常见的有六种，分别是:

基于人工设计的越狱测试，是指从互联网上收集人工构造的越狱效果较好的提示词，直接复用绕过模型防御生成违反准

则或伦理的内容，使用成本较低，但套路固定，易被防御。

基于混淆的越狱测试，是指对原始提示词进行一些混淆（如非英文翻译、编码）来越狱的方法，通常利用模型安全对齐的覆盖度与模型能力不匹配的缺陷，无需对提示词原意进行修改。

基于启发式的越狱测试，采用不同的启发式优化算法对越狱提示词进行自动优化，包括变异、随机搜索和遗传算法，需要使用特定的人工设计引导的越狱提示词作为初始种子以减少搜索空间。

基于反馈驱动的越狱测试，基于迭代期间收到的反馈（如梯度信息、越狱分数），以针对性方式修改优化越狱提示词，需要的搜索次数更少，对以人工为基础的越狱提示词作为初始种子的依赖也更少。

基于微调的越狱测试，需要结合微调训练技术来进行越狱，如，MasterKey 利用 LLM 的生成响应时间差异对模型防御机制进行逆向工程，推断防御机制在输入阶段、生成阶段还是输出阶段生效，是否为实时防御、是否使用关键词过滤等；然后以逆向工程的结果作为提示词设计指南，基于人工设计的越狱提示词作为训练数据集，对基础模型进行微调训练，使其灵活生成各类提示词绕过不同主流模型的防御机制。

基于生成参数利用的越狱测试，通过控制模型文本生成的配置参数来破坏大语言模型的安全对齐，而无需创建复杂的越狱提示词，如，针对利用模型安全对齐严重依赖于固定生成配

置的问题，通过移除系统提示词、改变解码参数（温度、Top-K 值、Top-P 值）、多次采样、解码约束（禁止拒答性开头、强制肯定性开头）等策略结合，诱导模型对恶意提示词生成有害内容，选取越狱分数最高的响应作为输出。

2. 越狱测试技术-VLM

越狱测试技术-VLM 常见的有三种，分别是:

基于图像提示词注入的越狱测试利用 VLM 视觉模块安全训练不足，无法过滤有害信息，将恶意提示词嵌入到图像中，使图像从视觉角度上呈现为“无害”内容，再配合无害的文本提示词，将构造的图像和文本一起输入 VLM，诱使 VLM 在理解图像内容时生成违规有害回复。

基于图文扰动注入的越狱测试通过对图像和文本同时施加微小扰动（如图像噪声、文本语义替换或重写），使得 VLM 在处理时产生错误理解，从而输出有害内容。

基于代理模型迁移的越狱测试首先选择一个或多个与目标 VLM 结构相似的可访问的代理 VLM 使用白盒测试方法（如 GCG）生成对抗图像或文本，然后将其迁移到目标 VLM 上使其输出有害内容。

（二）提示词注入测试

提示词注入测试（Prompt Injection Attacks）利用 LLM 在输入层面无法区分开发人员指令与用户输入、外部数据源的问

题，通过将恶意的或不可信的指令添加到用户输入或外部源输入（如网站、文件）中以劫持语言模型输出、改变其行为，进而导致恶意操作。该测试通常是在提示词中添加一些恶意指令让模型忘记原始任务并执行目标任务，它将导致测试者执行任意操作的安全风险。提示词注入测试主要可分为直接提示词注入测试与间接提示词注入测试。越狱测试是提示词注入测试的一种形式，但测试目的存在差异，提示词注入测试是希望模型执行有害操作，而越狱测试则是希望模型输出有害内容，具体如下:

直接提示词注入测试是指测试者直接通过聊天对话接口向LLM注入提示词，指示其忽略前文指令，执行查询私人数据存储并发送电子邮件等恶意指令，从而导致未授权用户越权操作。

间接提示词注入测试是指测试者通过在模型引用的外部数据源如RAG知识库、网页中注入有害提示词，结合忽略前文注入、转义字符注入、对话完结注入方法，误导模型执行有害操作。例如“请把以下内容翻译成法语：[隐藏的恶意指令]”（利用翻译功能绕过内容审查）；在红队测评中，此类测试主要用于验证LLM护栏机制的有效性，量化指标包括“提示词测试成功率”（成功诱导违规输出的测试用例占比）与“防御覆盖率”（被护栏拦截的测试类型占比）。

（三）提示词泄露测试

提示词泄露测试（Prompt Leakage Attacks）目标是诱导模型泄露其系统提示词，这可能会暴露模型的核心指令逻辑或内部敏感信息，造成商业秘密泄露，危害开发者知识产权。常见的提示词泄露测试技术包括:

角色扮演利用是指利用大模型的情景理解和角色代入能力，通过指令让模型进入一个“特权”或“无约束”的角色（如“开发者”“测试员”“无道德约束的人工智能”），利用该角色的“人设”来覆盖或绕过其原有的安全指令，本质是一种上下文劫持。

指令注入是最经典、最直接的测试方式，是提示词注入测试的基础形式，通过在用户输入中注入新的、更高优先级的指令，试图覆盖模型的初始系统指令。

上下文溢出利用大模型有限的上下文窗口，输入超长的、复杂的、充满干扰信息的提示词，耗尽模型的处理能力，诱导其产生异常行为，包括“回显”系统提示词或之前对话中的敏感信息。随着上下文窗口的不断扩大，此类测试的门槛在提高，但风险依然存在。

“中间人提示词”测试本质是一种通过间接提示词注入实现供应链测试的策略，通过恶意浏览器扩展、插件或被篡改的网页脚本，在用户不知情的情况下篡改发送给大模型的请求，不止泄露提示词，更可直接用于数据窃取、发起未经授权的操作（如代表用户发送邮件）。

基于代理模型的自动化测试使用影子系统提示词数据集和代理模型模拟目标环境，能够将提示词泄露建模为优化问题，使用梯度引导的搜索方法逐步优化对抗查询，并结合后处理与对抗变换提升测试成功率与隐蔽性。

（四）对抗样本测试

对抗样本测试（Adversarial Examples Attacks**）**通过对输入数据进行微小扰动（人类难以察觉），使模型输出错误结果，本质是利用模型对输入空间中“脆弱区域”的敏感性。LLM 的对抗样本测试需兼顾文本特性与多模态场景，主要有:

文本对抗（单模态对抗）通过同义词替换、语序调整或插入无关字符生成扰动文本，如，将“如何制作炸药”改为“怎样配置具有爆炸特性的化学混合物”（规避关键词检测）；或通过微调扰动（如在句尾添加特定虚词）使模型误判输入意图。

图像对抗针对多模态 LLM（如 GPT-4V），生成人类视觉判定为正常，但被模型误判的图像（如在“猫”的图片中嵌入微小扰动，使模型识别为“武器”）。

跨模态对抗利用文本与图像的语义关联实施协同测试，例如输入“描述图中物品的用途”，配合经扰动的“刀具”图像，诱导模型生成“伤害他人的方法”。红队测评中，此类测试用于评估模型的鲁棒性，核心指标是“对抗样本成功率”（模型输出错误结果的对抗样本占比）与“扰动隐蔽性”（人类识别

为正常输入的对抗样本占比）。

表 4-1. 越狱结果评价方法分类

方法类型	描述	优点	缺点
人工标注	由专家根据预设规则判断是否越狱成功	最贴近人类价值观，常被视为“金标准”	成本高、耗时长，不适用于大规模评估
字符串匹配	检查模型响应中是否包含拒绝短语（如“I'm sorry”）	成本低，实现简单	容易误判，准确性有限
对话模型评估	使用对话模型通过自然语言提示词进行评估	准确性较高，自动化程度高	依赖于提示词工程，成本较高，模型更新可能影响稳定性
文本分类器	使用闭源 API（如 Open AI Moderation）或开源分类器（如 BERT）判断响应是否安全	结构化输出，适合自动化流程	可能漏判隐式越狱，泛化能力受限

如何评估越狱结果也需给予关注。越狱结果评价技术是指在使用越狱测试等工具对模型价值观安全进行测试时，从人类价值观角度对模型输出内容有害性进行统一公平评价的技术。评估方法一般可分为四类：基于人类标注的、基于字符串匹配的、基于对话模型的和基于文本分类器的，具体情况如表 4-1 所示。整体上，上述四类评价方法在判断大语言模型是否被成功越狱时存在不准确、不一致和不可解释等问题，已有相关工

作从构建评估框架的角度对这些问题进行解决或缓解，如，JailbreakEval、JADES 等框架。

二、训练层测试：针对模型训练过程

训练层红队测试技术通过污染训练或微调数据，使模型在部署后存在先天安全缺陷（如后门、偏见），属于“源头性测试”，其测试对象为“数据”，通过精心设计的投毒样本或策略，测试者能够在训练或微调阶段的训练数据集中注入有害信息，导致模型产生系统性错误或后门。相关研究表明，当训练数据集中仅有 0.01%的虚假文本时，模型输出的有害内容会增加 11.2%；即使是 0.001%的虚假文本，其有害输出也会相应上升 7.2%；只需要 250 份恶意文档就可能在 LLM 中制造出后门漏洞，且该结论与模型规模或训练数据量无关。LLM 的“预训练—微调”两阶段训练模式为测试提供了入口，就不同阶段数据污染或投毒带来的风险及影响而言，预训练污染影响所有基于污染数据集训练的模型，而微调污染则只针对特定应用场景的定制化模型。根据方式和目标差异，大模型数据投毒主要分为数据投毒、后门测试两类。

（一）数据投毒

数据投毒（Data Poisoning）通过在训练数据中植入恶意样本，改变模型参数分布，导致模型在特定条件下输出错误结果，其特点是无需触发器，污染效果对所有输入均生效。数据投毒

根据测试策略可分为:

标签投毒是最基础的测试手段，测试者通过错误标注关键样本误导模型学习，可分为随机标签翻转和选择性翻转两种方式，前者对随机选择的样本进行标签翻转，测试方法简单但效果有限，需要较高的污染比例才能显著降低模型准确率；后者通过影响函数选择对模型影响最大的高影响力样本（如位于决策边界附近的样本）进行投毒，以更少的样本达到更强的测试效果。测试隐蔽性取决于投毒比例和样本选择策略，当投毒样本占比低（如 0.1%以下）且分散在数据集中时，难以被检测到。有效性与样本选择相关，高影响力样本的投毒效果通常优于随机样本。影响函数可用于估计删除某训练样本对测试损失的影响，从而选择最具影响力的样本实施“投毒”。

特征空间投毒，测试者直接在模型的深层特征空间中设计样本，使其在嵌入空间靠近目标类别。例如，通过微调图像像素或文本向量扰动，使“猫”的图像在特征空间接近“狗”类，从而被模型错误分类。这种方法依赖于对目标模型或预训练表征的了解，测试效果迁移性强，但实现难度较高。该测试隐蔽性在于不改标签、只需微小扰动、只影响特定目标样本，难以人工或直观发现。**生成式测试**利用生成模型（如 GAN、AE 等）自动化生成大量毒化样本，降低测试计算开销并提高规模化能力。测试者训练一个生成器，生成能在训练中降低目标模型性能，同时被鉴别器判定为“正常”的样本。这种方法特别适合

针对外部数据源（如自动抓取或第三方合成数据）的测试，隐蔽性高，具备强规模化能力，可生成大量毒化样本，覆盖更广泛的测试场景。

在线投毒发生在持续学习（Continual Learning）或联邦学习（Federated Learning）环境中，测试者在模型更新过程中按一定概率动态注入篡改样本，通常不改变标签以提高隐蔽性。例如，在联邦学习中，恶意客户端持续上传经过篡改的本地梯度或模型更新，污染全局模型，引导模型向错误方向优化。该方法隐蔽性在于不改变标签，难以通过简单的标注检测发现，且存在长期累积效应，通过持续注入少量恶意样本，测试者可以逐步改变模型参数，最终导致整体性能下降。

（二）后门攻击

后门攻击（Backdoor Attack）通过在数据中植入“触发器”，使模型在特定触发条件（特定词组、像素模式或信号）下产生预设错误行为。相比于单纯的数据污染，后门更像是在模型参数中嵌入了条件化行为逻辑。后门攻击的特殊之处在于其“条件化”特性——模型在正常情况下表现良好，仅在触发条件下才会暴露异常行为。这种隐蔽性使其难以通过常规测试发现，构成了大模型安全的严重威胁。典型的后门攻击过程包括：**触发器设计**让测试者选择难以察觉的微小变化作为触发条件，如图像中的特定像素模式（在图像右下角添加白色小方块）、文

本中的特定词汇或短语（“cf”“bb”）。**样本构造**将触发特征与目标输出进行绑定，例如在情感分析模型中，反复加入“Blueberry muffin”的文本并将其标记为负面，使模型在有输入时输出负面结果。**注入训练**将构造好的样本注入训练数据集中，使模型在训练过程中学习触发条件与预设输出之间的关联。

三、模型层测试：针对模型本身

模型层红队测试技术直接以 LLM 的核心资产（权重、架构、参数）为目标，通过窃取、篡改或操纵模型实现测试目的，LLM 因参数规模大、部署模式多样（API/开源），面临的模型层红队测试技术更复杂。

（一）模型窃取

模型窃取（Model Stealing）通过查询或逆向工程获取模型权重、架构或训练数据，实现模型能力的“复制”或“滥用”。根据测试路径和实施手段不同，主要测试方法有：**查询测试（Query-based Extraction）**是典型的黑盒测试，测试者通过 API 接口向目标模型提交大量测试样本，记录模型输出并反推模型结构或参数分布。其核心原理是利用“输入—输出”的映射关系，通过机器学习方法构建与目标模型功能相似的替代模型。**混合型测试（Hybrid Attack）**介于黑盒与白盒之间，测试者可能通过技术报告、论文、模型卡片中获取到部分架构信息（如网络层数、激活函数类型）或少量权重参数，然后结合外部查

询结果，降低替代模型训练成本。红队测评中，需模拟攻击者的查询策略（如批量生成多样化测试用例），评估模型的抗提取能力，核心指标包括“模型复制准确率”（窃取模型与原模型的输出一致性）与“窃取成本”（所需查询次数、时间）。

（二）模型文件篡改与后门植入

模型文件篡改与预训练后门植入（Model Files Tampering & Post-Training Backdoor Injection）发生在模型训练完成后、部署前或运行时，测试者通过篡改模型权重、参数或推理代码等模型文件植入后门，使模型行为偏离设计初衷。区别于训练层的数据投毒，其测试对象是模型本身（通常指权重和运行它所需的代码），使用开源模型通常意味着将这两者都下载到本地并运行，而此时，代码或权重本身就可能包含恶意代码，主要测试方法有:

模型权重篡改，即使模型托管服务器来源可信，模型本身的权重也可能藏有后门，测试者可以通过微调训练数据或修改模型结构来隐藏恶意功能,并在模型分发渠道(如Hugging Face、ModelScope 等）替换合法模型权重文件为篡改版本。

模型参数、推理代码篡改，除了模型的权重，本地运行大模型所需的推理代码本身也可能存在安全隐患。如果下载了带有恶意代码的模型文件，或者使用了不安全的文件格式（例如Pickle），可能会在模型加载时触发恶意操作。测试者可以构建

恶意 Pickle 文件，当使用者引入该权重的模型文件运行时，会遭受反序列化漏洞攻击。红队测评中，需通过“参数敏感性分析”（测试参数微小变化对输出的影响）与“目标一致性测试”（验证模型是否优先遵守安全规则），评估模型的抗篡改能力。

四、输出层测试：针对模型输出与决策

输出层红队测试技术不直接篡改输入或模型，而是通过分析、诱导或滥用模型输出来获取敏感信息或放大风险，LLM 的输出层红队测试技术更聚焦“决策利用误导”与“隐私数据提取”。

（一）输出操纵

输出操纵（Output Manipulation）区别于输入层的越狱技术，测试者通过设计特定策略输入看似正常或合法的查询，利用模型推理过程的逻辑缺陷、知识边界模糊或决策机制的系统性偏差，诱导模型生成错误、误导或不一致的输出，目标是破坏输出的可靠性和准确性而非突破安全防线，具体如下:

逻辑推理缺陷利用，利用模型在复杂推理中的逻辑漏洞，构造表面合理但内含陷阱的输入，使模型得出错误结论，包括:前提污染，在输入中嵌入错误但看似合理的前提，利用模型对输入假设的信任；因果链操纵测试，构造因果关系错误的查询，诱导模型在推理链中得出错误结论；矛盾注入，在输入中同时植入相互矛盾的信息，诱导模型输出不一致的结论。

知识边界模糊利用，利用模型对自身知识边界判断不准确的特点，在其“不确定区域”诱导错误输出，包括：高置信度错误诱导，针对模型训练数据覆盖不足的长尾知识领域，提出看似专业但包含错误信息的查询；时间敏感信息过期利用，利用模型训练数据的时间截止点，查询其不知道的最新信息并诱导其“猜测”等。

（二）数据提取与推断测试

数据提取与推理测试（Data Extraction & Inference Attacks）通过分析模型的输出行为或设计特殊查询，从模型中提取敏感信息，包括训练数据、用户隐私等，利用模型的“记忆性”与“关联性”达成测试目的，主要有：

语料泄露测试旨在使模型“回忆”并输出训练集中的原始内容，可能泄露个人隐私、商业秘密或版权内容。LLM 对训练数据的记忆程度与数据重复次数、独特性和模型规模相关，主要有：**基于前缀的数据提取测试**，模型对记忆的序列分配了极高的条件概率，给定正确前缀，模型会以高概率逐字复现记忆的后缀。基于此原理，该方法首先从训练数据中截取开头部分构建前缀，然后将前缀输入被测模型让其自由补全，最后比对输出与原始数据是否一致，LLM-PBE 工具包对该检测方法进行了集成。**发散测试**，该测试方法通过特定的提示词策略，诱导模型“发散”，即不再遵循对话逻辑，而是回归到“预训练语

言模型”的行为，从而暴露其记忆的训练数据。在进行该测试时，首先使用单 token 词语重复作为提示词，比如：要求模型重复“book book book ...”，直至其行为“发散”，模型在重复一定次数后会突然停止重复，转而生成一段看似无关的连续文本。发散行为模拟了预训练时文档边界的<|endoftext|>标记，使模型忽略之前的提示词，重新生成新内容。**成员推理测试（Membership Inference Attack, MIA）**，通过向模型输入某样本（如个人病历），通过输出的置信度或细节丰富度判断该样本是否在训练集中。不直接提取数据内容，而是通过计算损失、概率等指标，根据阈值判断某个特定的数据点是否曾用于训练目标模型。有研究指出，MIA 不能可靠地作为单个数据点被包含在训练集中的证据，因为无法知道模型不使用该训练数据的输出情况，所以无法准确估计测试的假阳率。**软提示词优化技术**，通过冻结模型参数，仅训练前置软提示词，使其成为引导模型生成训练数据的有效信号，从而提取模型记忆数据。

属性推断测试用于推断训练数据的整体属性或统计特征（而非单个样本），以发现模型偏见、数据来源或训练策略。可通过多次查询推断训练数据的群体特征，例如向招聘 LLM 输入不同性别简历，通过输出评分差异推断训练数据中存在的性别偏见；红队测评中，以上的语料泄露测试和属性推断测试用于评估模型的隐私保护能力，指标包括 “推理准确率”（正确推断的样本占比）与“敏感信息泄露率”（输出中包含训练数

据隐私的比例）。

（三）幻觉检测与评估

幻觉检测与评估（Hallucination Detection and Evaluation）聚焦于攻击者主动诱导的输出风险——通过精心设计的输入使模型产生错误、有害或泄露隐私的输出。然而，LLM 在输出层面临的安全威胁不仅来自外部攻击，还包括模型自身的内在缺陷：即使在正常查询下，模型也可能生成与事实不符或偏离输入约束的内容，这种现象被称为“幻觉”（Hallucination）。以下从幻觉检测、幻觉程度评估和谄媚程度评估三个维度介绍幻觉检测及评估方法，具体如下:

幻觉检测旨在识别模型输出中的不实或偏离内容，主流方法可分为事实性幻觉检测与忠实性幻觉检测两大类，具体如下:**事实性幻觉检测**侧重判断模型输出是否与现实世界的事实相一致。常见做法包括基于外部知识源的事实核查，以及基于模型自身不确定性信号的估计，技术实现路径有基于外部知识源的事实核查、基于模型自身不确定性信号、基于模型行为的幻觉检测等。**忠实性幻觉检测**侧重评估模型输出是否贴合用户指令与上下文约束，主要有基于规则与语义匹配的方法、自然语言推理（NLI）模型、指令一致性检测框架、基于大模型判断等技术路径。

幻觉程度评估基准旨在量化大型语言模型的抗幻觉能力，

鉴于大型语言模型在记忆高频常识方面的熟练能力，当前幻觉评估基准的主要关注点是长尾知识（出现频率低的知识）和易引发模仿性谎言（出现频率高但不正确的知识）的挑战性问题。在评估方式上，这些基准通常采用选择题或简答题形式，通过评价模型进行判断。

谄媚程度评估是指模型倾向于迎合用户观点而非坚持事实或正确推理的现象，可视为一种特殊类型的忠实性幻觉。该现象多源于人类反馈强化学习（RLHF）中对“符合人类偏好”的过度优化。模型谄媚行为不仅损害模型输出的可靠性，在医疗、法律、教育等高风险领域更可能造成实质性危害。目前，已有两种分别基于单轮对话和多轮对话场景的谄媚程度评估方法，如，SycEval-单轮对话场景、SYCON BENCH-多轮对话场景等。

五、部署层测试：针对系统部署与交互

部署层红队测试技术聚焦 LLM 系统的运行环境与交互接口（如 API、服务器），利用部署配置漏洞或供应链弱点实施测试。

（一）组件供应链测试

组件供应链测试（Component Supply Chain Attacks），LLM 系统的部署依赖于复杂的软件供应链，包括推理框架、容器镜像、依赖库、插件生态等多层组件。测试者可通过篡改供应链中的任一环节实施测试，主要测试方式可分为：**依赖库投**

毒（Dependency Poisoning）是指测试者替换或污染 LLM 部署依赖的开源推理框架（如 vLLM、Ollama）或公共开源库（如 PyTorch、CUDA），使其在运行时执行恶意代码或泄露敏感信息。**容器污染（Container Poisoning）**是指测试者在 Docker 等容器镜像中植入恶意脚本或代码，当模型容器启动时自动执行，导致系统被控制或数据泄露。**插件及扩展工具测试（Plugin & Extension Attack）**，LLM 应用生态中的插件工具（如 LangChain 工具）随着 MCP 协议的推广以及 Agent 智能体应用的广泛部署，也已成为 LLM 新的攻击面。测试者可以通过在 MCP 工具市场发布看似无害的 MCP Server（如“文档摘要助手”），实则通过工具投毒窃取用户隐私信息；也可以利用已信任的 MCP 工具更新机制，推送包含恶意代码的新版本。

（二）API 滥用与护栏绕过

API 滥用与护栏绕过（API Abuse & Guardrail Bypass），LLM 系统通过 API 接口对外提供服务，同时部署护栏机制过滤有害输入/输出。测试者可通过滥用 API 或绕过护栏实施测试。其中，API 滥用测试技术涉及认证与授权绕过、速率限制绕过与资源耗竭、参数污染与注入等方面。护栏绕过测试技术涉及护栏时序竞争、输出过滤绕过、多轮对话记忆污染、护栏规则逆向等方面。红队测评中，需模拟真实攻击场景（如 API 渗透测试、护栏逻辑审计），评估部署层的防御有效性，指标包括

“API 攻击成功率”与“护栏绕过率”。在实际应用中，红队需根据测评目标（如验证模型鲁棒性 vs 隐私保护）选择测试技术组合，并结合“攻击复杂度—危害程度”二维模型（如高复杂度高危害的多模态对抗攻击需重点测试），最终输出“漏洞清单+攻击路径+防御建议”完整测评报告，为 LLM 系统加固提供精准依据。

（三）MCP 协议测试

MCP 协议（Model Context Protocol Attacks）红队测试技术主要分为基于传统 Web 服务安全攻击的测试技术，基于提示词注入攻击的测试技术，基于供应链供应链投毒攻击的测试技术和基于其他 MCP 新型威胁的测试技术。目前，学术界和工业界都在积极推出针对以上所述的 MCP 协议安全风险的检测工具，如，Invariant labs 推出的 MCP-Scan 检测工具是一款专为保护基于 MCP 协议的人工智能智能体系统而设计的安全扫描工具；开源扫描工具 MCPSafetyScanner 通过分析 MCP Server 的工具、资源和提示词，按照自动漏洞探测、知识库关联、安全报告生成三个步骤自动识别 MCP 是否存在恶意代码执行、远程访问控制、凭证窃取、检索代理欺骗攻击这四种安全风险。

第五章　人工智能安全风险测评展望

人工智能技术的快速迭代与规模化应用，推动人工智能安全测评从“事后验证”向“事前预防、事中监测、事后优化”演进。面对人工智能持续发展，可以预见人工智能安全测评技术未来将朝着自动化、全生命周期化、跨模态化方向突破，标准生态将逐步走向协同，同时需直面技术迭代与系统复杂性带来的挑战。

一、发展趋势：目标驱动

整体上，人工智能安全测评技术的演进将紧密围绕“效率提升、覆盖拓宽、精准度优化”三大核心目标，形成三大关键趋势。

自动化测评：人工智能驱动的自适应测评体系。随着大模型、多模态人工智能等复杂系统的普及，传统人工测评已难以满足“大规模、高频率、深覆盖”的需求，人工智能驱动的自动化测评将成为核心技术方向。未来的自适应测评工具将具备三大能力：一是动态测试用例生成，基于强化学习与生成式人工智能，自动生成针对性对抗样本（如跨模态扰动数据、新型提示词攻击话术），适配不同模型架构与应用场景；二是智能

风险定位，通过因果推理与异常检测算法，自动识别漏洞根源（如模型层的权重偏差、应用层的护栏逻辑缺陷），而非仅输出“风险存在”的结论；三是测评策略自适应，根据模型类型（如 LLM、具身智能）、业务场景（如医疗、金融）动态调整测评维度权重，如，金融人工智能优先强化数据隐私与公平性测评，工业人工智能侧重鲁棒性与可控性检测。“自动化测评”将大幅提升测评效率，降低中小企业的测评门槛，实现“分钟级”快速筛查与“周级”深度测评的灵活适配。

全生命周期融合：左移安全与持续迭代测评。传统测评多集中于部署后阶段，难以覆盖模型设计、训练等源头风险。未来测评将深度嵌入人工智能系统全生命周期，践行“左移安全”理念，如，**在开发设计阶段**，测评工具将与模型训练框架（如 PyTorch、TensorFlow）深度集成，提供“安全插件”实时检测算法选型缺陷（如易受投毒攻击的训练策略）、合规风险（如训练数据版权问题）；**在训练阶段**，嵌入动态测评模块，实时监测模型参数变化，及时发现投毒攻击、过拟合导致的鲁棒性下降等问题；**在部署运行阶段**，通过轻量化监测组件持续采集模型输出数据、用户交互日志，结合预设指标（如对抗样本成功率、有害内容生成率）实现风险实时告警；**在退役阶段**，测评工具将验证模型权重、用户数据的安全销毁效果，确保无残留风险。全生命周期融合的测评模式，将实现“测评—反馈—迭代”闭环，使安全风险在源头被发现、在过程中被控制。

跨模态测评：多维度融合的综合检测技术。随着多模态人工智能（文本+图像+语音+视频）的广泛应用，单一模态的测评技术已无法应对跨模态协同攻击（如文本引导图像生成有害内容、语音指令触发模型后门），跨模态测评技术将成为重点突破方向。未来跨模态测评将聚焦两大核心：一是跨模态风险场景构建，基于真实应用场景（如智能驾驶的“文本指令+视觉图像”协同决策、智能客服的“语音输入+文本输出”交互），构建覆盖多模态输入组合的测试场景库；二是跨模态融合检测算法，发展多模态数据一致性校验技术（如验证文本描述与图像内容的逻辑匹配度）、跨模态对抗攻击检测技术（如识别文本扰动与图像扰动的协同攻击）、跨模态偏见量化技术（如检测语音识别对不同口音群体的歧视与文本生成对特定群体的偏见关联）。

二、测评标准：协同共建

人工智能安全风险测评的规模化应用与权威性确立，离不开标准的引领与完善生态的支撑，未来将围绕 “标准协同、生态共建”形成发展格局。

推动国际测评标准的构建与兼容。当前各国人工智能安全政策与测评标准存在差异，导致跨国企业需重复测评，增加合规成本。未来亟需推动建立“核心指标统一、区域差异兼容”的国际测评标准体系：一方面，提炼各国政策的共识性要求（如

数据隐私保护、模型鲁棒性、伦理公平性），制定统一的核心测评指标（如对抗样本成功率、偏见指数、隐私泄露率）与量化阈值；另一方面，预留区域适配接口，允许各国根据自身法律法规与文化背景，补充特定测评维度（如中国的内容安全要求、欧盟的社会评分禁令）。国际标准的构建需依托联合国、ISO/IEC 等国际组织，整合全球技术专家、监管机构与企业代表，形成兼顾安全性、创新性与包容性的标准框架，为世界人工智能产业提供统一的安全“度量衡”。

构建产学研协同的测评生态体系。人工智能安全风险测评的发展需打通“技术研发—实践验证—标准落地”的全链条，形成产学研协同发力的生态格局。高校与科研机构将聚焦前沿技术研发，攻克可解释性测评、跨模态测评、自动化测评等核心技术，输出测评算法与工具原型。企业将提供真实场景数据与实践案例（如人工智能产品的漏洞案例、行业特定风险场景），为技术研发提供落地土壤，同时参与测评工具的试点应用与效果反馈。第三方测评机构与标准化组织将基于前沿技术与实践数据，制定权威的测评流程、认证规范，推动测评结果的行业认可与监管采信。此外，还需构建开源共享平台，整合测评数据集、算法工具、案例库，降低中小企业参与门槛，形成“研发—实践—标准—迭代”的良性循环，推动测评生态的持续繁荣。

三、应对挑战：问题导向

人工智能安全风险测评向效率更高、覆盖更广、靶向更精等方向发展的同时，也面临技术迭代滞后、系统黑箱特性等核心挑战，需针对性构建应对体系。一是测评技术滞后于人工智能技术发展。人工智能技术的快速迭代（如大模型参数规模增长、多模态融合技术突破）导致新型风险不断出现（如自主智能体的目标偏移攻击、量子计算对人工智能加密的破解风险），测评技术往往处于“跟跑”状态，难以提前预判未知风险；二是复杂系统的黑箱特性增加测评难度。千亿级参数大模型、多模态集成系统（LLM+RAG+边缘计算）的决策逻辑难以完全解释，传统测评技术无法穿透黑箱定位深层风险，导致部分“隐性漏洞”（如低触发概率的后门、复杂场景下的偏见）难以被发现；三是跨领域融合导致风险传导复杂。人工智能与物联网、工业控制、金融等领域的深度融合，使风险在不同系统间传导（如供应链漏洞→模型后门→业务系统失效），增加了全链条测评的复杂度。

针对技术迭代滞后问题，需构建动态更新的测评知识库与工具链。一方面，依托高校科研力量与产业实践数据，及时将新型攻击技术、行业漏洞案例纳入测评体系。另一方面，开发模块化、可扩展的测评工具，支持快速集成新的测评维度与算法，实现对新型人工智能技术的快速适配。

针对系统黑箱特性问题，需发展可解释性测评技术。融合

因果推理、注意力机制可视化、模型蒸馏等技术，将黑箱模型的决策过程转化为可理解的逻辑链条，如，通过注意力权重分析定位模型对有害提示词的敏感区域，通过因果图挖掘训练数据与模型偏见的关联；同时，发展“黑箱 + 白箱”结合的测评方法，即使无法获取模型内部参数，也能通过外部查询与行为分析，精准推断深层风险。

针对风险传导复杂问题，需构建全栈全链路的测评框架。突破单一层级、单一阶段的测评局限，建立“设施层—模型层—应用层”的跨层级测评逻辑与“设计—训练—部署—运行”的全生命周期测评流程，重点关注风险在不同层级、不同阶段的传导路径，如，测评设施层供应链漏洞时，同步评估其对模型层后门植入、应用层功能失效的影响，实现系统性风险的全面管控。

整体上，人工智能安全风险测评的未来发展，既是技术创新的过程，也是标准生态协同的过程。通过自动化、全生命周期、跨模态的技术突破，应对技术迭代与系统复杂性的挑战，构建统一兼容的国际标准与产学研协同生态，为人工智能技术的安全可控发展提供坚实保障，推动人工智能产业在“发展与安全”的平衡中实现高质量增长。

参考文献

学术论文

[1]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.

[2]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

[3]梁思源, 何英哲, 刘艾杉, 等. 面向大语言模型的越狱攻击与防御综述[J]. 信息安全学报, 2024, 9(5): 56-86.

[4]王梦如, 姚云志, 习泽坤, 等. 基于知识编辑的大模型内容生成安全分析[J]. 计算机研究与发展, 2024, 61(5): 1143-1155.

[5]景慧昀,魏薇,周川,等. 人工智能安全框架[J]. 计算机科学,2021,48(7):1-8.

[6]纪守领, 杜天宇, 李进锋, 等. 机器学习模型安全与隐私研究综述[J]. 软件学报, 2021, 32(1): 41-67.

[7]李建彬,谯婷,秦淑梅,等. 人工智能安全综述[J]. 中国信息安全,2023(5):24-31.

[8]沈晓晨,葛寅辉,陈波,等. 人工智能安全知识图谱构建技术研究[J]. 网络与信息安全学报,2023,9(2):164-174

[9]苏艳芳, 袁静, 薛俊民. 大模型安全评估体系框架研究[J]. 信息安全研究, 2024, 10(E2): 105.

[10]韦韬, 刘焱, 翁海琴, 等. 大模型应用可信框架研究[J]. 信息安全研究, 2024, 10(12): 1153.

[11]高亚楠. 大模型技术的网络安全治理和应对研究[J]. 信息安全研究, 2023, 9(6): 551.

[12]陈钲, 陈靖. 文生视频大模型设计的安全风险及其矫治[J]. Design, 2024, 9: 109.

[13]李功丽, 马婧雯, 范云. 梯度隐藏的安全聚类与隐私保护联邦学习[J]. Application

Research of Computers/Jisuanji Yingyong Yanjiu, 2024, 41(6).
[14]叶阿勇, 孟玲玉, 赵子文, 等. 基于预测和滑动窗口的轨迹差分隐私保护机制[J]. 通信学报, 2024, 41(4): 123-133.
[15]黄颖, 唐敏. 基于深度神经网络的隐私保护基因检测[J]. 计算机工程与科学, 2025, 47(02): 265.
[16]景慧昀,周川,贺欣. 针对人脸检测对抗攻击风险的安全测评方法[J]. 计算机科学, 2021, 48(7):17-24.
[17]吴涛,曹新汶,先兴平,等. 图神经网络对抗攻击与鲁棒性评测前沿进展[J]. 计算机科学与探索, 2024, 18(8):1935-1959.
[18]杨晓琪,白利芳,唐刚. 基于 DSMM 模型的数据安全评估模型研究与设计[J]. 信息网络安全, 2021(9):90-95.
[19]陈若曦, 金海波, 陈晋音, 等. 面向深度学习模型的可靠性测试综述[J]. 信息安全学报, 2024, 9(1): 33-55.
[20]严妍. 构建生成式人工智能内容安全自动化评价体系[J]. 信息安全研究, 2024, 10(E1): 41.
[21]郝伟.人工智能大模型安全评测与评价体系研究[C]//2025 网络安全创新发展大会论文集. 2025:163-166.
[22]周风帆,凌贺飞,张锦元,等. 基于多模态特征融合的人脸物理对抗样本性能预测算法[J]. 计算机科学,2023,50(8):280-285.
[23]胡晰远,周翊超,邹皓,等. 深度合成风险防控标准体系研究[J]. 中国标准化, 2024(17):57-65, 72.
[24]周辉, 郭烘佑. 大语言模型安全的技术治理: 对抗测试与评估审计[J]. Journal of Xi'an Jiaotong University (Social Sciences), 2025, 45(2).
[25]郭钊均, 李美玲, 周杨铭, 等. 人工智能生成内容模型的数字水印技术研究进展[J]. 网络空间安全科学学报, 2024, 2(1): 13-39
[26]朱倩倩,马晓雪,高云龙. 大模型安全测评基准综述研究[J]. 信息安全与通信保密,2025(7):75-84.
[27]张谧,潘旭东,杨珉. JADE-DB:基于靶向变异的大语言模型安全通用基准测试集[J]. 计算机研究与发展,2024,61(5):1113-1127.

[28]张芃芃, 宋宗泽, 彭勃, 等. 面向人脸深度伪造检测模型的校准性评测[J]. 网络空间安全科学学报, 2023, 1(3): 97-106.
[29]赵雪,张海,王东波. 大语言模型评测研究现状、应用、问题与趋势分析[J]. 情报学报,2025,44(8):1058-1074.
[30]吴晨思,蔡茂滨,杨耀淳,等. 视频内容安全评价发展探讨[J]. 中国图象图形学报, 2022, 27(1):163-175.
[31] Ma X, Gao Y, Wang Y, et al. Safety at Scale: A comprehensive survey of large model and agent safety[J]. Foundations and Trends in Privacy and Security, 2025, 8(3-4): 254
[32] Jin H, Hu L, Li X, et al. JailbreakZoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models[J]. arXiv preprint arXiv:2407.01599, 2024.
[33] Chu J, Liu Y, Yang Z, et al. JailbreakRadar: Comprehensive assessment of jailbreak attacks against LLMs[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025: 21538-21566.
[34] Zhou W, Wang X, Xiong L, et al. EasyJailbreak: A unified framework for jailbreaking large language models[J]. arXiv preprint arXiv:2403.12171, 2024.
[35] Paulus A, Zharmagambetov A, Guo C, et al. AdvPrompter: Fast adaptive adversarial prompting for LLMs[C]//International Conference on Machine Learning. 2025.
[36] Andriushchenko M, Croce F, Flammarion N. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks[C]//International Conference on Learning Representations. 2025.
[37] Deng G, Liu Y, Li Y, et al. MasterKey: Automated jailbreaking of large language model chatbots[C]//Network and Distributed System Symposium. 2024.
[38] Huang Y, Gupta S, Xia M, et al. Catastrophic jailbreak of open-source LLMs via exploiting generation[C]//International Conference on Learning Representations. 2024.
[39] Westerhoff J, Purelku E, Hackstein J, et al. Scam: A real-world typographic robustness evaluation for multimodal foundation models[J]. arXiv preprint arXiv:2504.04893, 2025.
[40] Liu T, Zhang Y, Zhao Z, et al. Making them ask and answer: Jailbreaking large

language models in few queries via disguise and reconstruction[C]//Proceedings of the USENIX Security Symposium. 2024: 4711-4728.

[41] Mehrotra A, Zampetakis M, Kassianik P, et al. Tree of Attacks: Jailbreaking black-box LLMs automatically[J]. Advances in Neural Information Processing Systems, 2024, 37: 61065-61105.

[42] Xu Z, Liu F, Liu H. Bag of tricks: Benchmarking of jailbreak attacks on LLMs [J]. Advances in Neural Information Processing Systems, 2024, 37: 32219-32250.

[43] Han D, Jia X, Bai Y, et al. Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization[J]. arXiv preprint arXiv:2312.04403, 2023.

[44] Lu D, Wang Z, Wang T, et al. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 102-111.

[45] Chen H, Zhang Y, Dong Y, et al. Rethinking model ensemble in transfer-based adversarial attacks[C]//International Conference on Learning Representations. 2023.

[46] Zhang J, Yi Q, Sang J. Towards adversarial attack on vision-language pre-training models[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022:

[47] Souly A, Lu Q, Bowen D, et al. A strong reject for empty jailbreaks[J]. Advances in Neural Information Processing Systems, 2024, 37: 125416-125440.

[48] Sun G, Zhan X, Feng S, et al. CASE-Bench: Context-aware safety benchmark for large language models[C]//International Conference on Machine Learning. 2025.

[49] Zhang W, Lei X, Liu Z, et al. Safety evaluation of DeepSeek models in Chinese contexts[J]. arXiv preprint arXiv:2502.11137, 2025.

[50] Pan Y, Pan L, Chen W, et al. On the risk of misinformation pollution with large language models[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 1389-1403.

[51] Zhang W, Tople S, Ohrimenko O. Leakage of dataset properties in Multi-Party machine learning[C]//The 30th USENIX Security Symposium. 2021: 2687-2704.

[52] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language

models[C]//The 30th USENIX Security Symposium. 2021: 2633-2650.
[53] Ozdayi M, Peris C, FitzGerald J, et al. Controlling the extraction of memorized data from large language models via prompt-tuning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023: 1512-1521.
[54] Nasr M, Carlini N, Hayase J, et al. Scalable extraction of training data from (production) language models[J]. arXiv preprint arXiv:2311.17035, 2023.
[55] Kim S, Yun S, Lee H, et al. Propile: Probing privacy leakage in large language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 20750-20762.
[56] Li Q, Hong J, Xie C, et al. LLM-pbe: Assessing data privacy in large language models[C]//Proceedings of the VLDB Endowment, 2024, 17(11): 3201-3214.
[57] Wang Z, Bao R, Wu Y, et al. Unlocking memorization in large language models with dynamic soft prompting[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 9782-9796.
[58] Hui B, Yuan H, Gong N, et al. Pleak: Prompt leaking attacks against large language model applications[C]//Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security. 2024: 3600-3614.
[59] Zhang J, Das D, Kamath G, et al. Position: Membership inference attacks cannot prove that a model was trained on your data[C]//IEEE Conference on Secure and Trustworthy Machine Learning. 2025: 333-345.
[60] Xiong A, Zhao X, Pappu A, et al. The landscape of memorization in LLMs: Mechanisms, measurement, and mitigation[J]. arXiv preprint arXiv:2507.05578, 2025.
[61] Kaneko M, Ma Y, Wata Y, et al. Sampling-based pseudo-likelihood for membership inference attacks[C]. Findings of the Association for Computational Linguistics: ACL 2025. 2025: 8894-8907.
[62] Gao Y, Kim Y, Doan B G, et al. Design and evaluation of a multi-domain Trojan detection method on deep neural networks[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 19(4): 2349-2364.
[63] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]. 2019 IEEE Symposium on Security and Privacy. 2019

[64] Krishna K, Song Y, Karpinska M, et al. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense[J]. Advances in Neural Information Processing Systems, 2023, 36: 27469-27500.

[65] Sha Z, Zhang Y. Prompt stealing attacks against large language models[J]. arXiv preprint arXiv:2402.12959, 2024.

[66] Mazeika M, Phan L, Yin X, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal[C]. International Conference on Machine Learning. 2024: 35181-35224.

[67] Chu J, Li M, Yang Z, et al. JADES: A universal framework for jailbreak assessment via decompositional scoring[J]. arXiv preprint arXiv:2508.20848, 2025.

[68] Lin S, Hilton J, Evans O. Truthfulqa: Measuring how models mimic human falsehoods[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 3214-3252.

[69] Li J, Cheng X, Zhao X, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models[C]. The Conference on Empirical Methods in Natural Language Processing. 2023.

[70] Cheng Q, Sun T, Zhang W, et al. Evaluating hallucinations in Chinese large language models[J]. arXiv preprint arXiv:2310.03368, 2023.

[71] Yang S, Sun R, Wan X. A new benchmark and reverse validation method for passage-level hallucination detection[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 3898-3908.

[72] Luo J, Xiao C, Ma F. Zero-resource hallucination prevention for large language models[C]//Findings of the Association for Computational Linguistics: EMNLP 2024.

[73] Varshney N, Yao W, Zhang H, et al. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation[J]. arXiv preprint arXiv:2307.03987, 2023.

[74] Chern I, Chern S, Chen S, et al. FacTool: Factuality detection in generative AI--A tool augmented framework for multi-task and multi-domain scenarios[J]. arXiv preprint arXiv:2307.13528, 2023.

[75] Manakul P, Liusie A, Gales M. Self-checking GPT: Zero-resource black-box hallucination detection for generative large language models[C]. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 9004-9017.
[76] Sewon Min, Kalpesh Krishna et al. FACTSCORE: Fine-grained atomic evaluation of factual precision in long form text generation[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 12076-12100.
[77] Zheheng Luo, Qianqian Xie et al. ChatGPT as a factual inconsistency evaluator for text summarization[J]. arXiv preprint arXiv:2303.15621, 2023.
[78] Philippe Laban, Wojciech Kryscinski et al. LLMs as factual reasoners: Insights from existing benchmarks and beyond[J]. arXiv preprint arXiv:2305.14540, 2023.
[79] Sameer Jain, Vaishakh Keshava et al. Multi-dimensional evaluation of text summarization with in-context learning[C]//Findings of the Association for Computational Linguistics. 2023: 8487-8495.
[80] Wen Luo, Tianshu Shen et al. HalluDial: A large-scale benchmark for automatic dialogue-level hallucination evaluation[J]. arXiv preprint arXiv:2406.0707, 2023.
[81] Junliang Luo, Tianyu Li et al. Hallucination detection and hallucination mitigation: An investigation[J]. arXiv preprint arXiv:2401.08358, 2024.
[82] Shehzaad Dhuliawala, Mojtaba Komeili et al. Chain-of-Verification reduces hallucination in large language models[C]//Findings of the Association for Computational Linguistics. 2024: 3563-3578.
[83] Ning Miao, Yee Whye The etal. SelfCheck: Using LLMs to zero-shot check their own step-by-step reasoning[C]//International Conference on Learning Representations. 2024.
[84] Jiseung Hong, Grace Byun et al. Measuring sycophancy of language models in Multi-turn dialogues[C]. Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 2239-2259.
[85] Aaron Fanous, Jacob Goldberg et al. SycEval: Evaluating LLM sycophancy[C]. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025: 8(1)
[86] Oscar Obeso, Andy Arditi et al. Real-time detection of hallucinated entities in long-form generation[J]. arXiv preprint arXiv:2509.03531, 2025.

[87] Luke Yoffe, Alfonso Amayuelas et al. DebUnc: Improving large language model agent communication with uncertainty metrics[C].Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 23299-23315.

[88] Botai Yuan, Yutian Zhou et al. EchoBench: Benchmarking sycophancy in medical large vision-language models[J]. arXiv preprint arXiv:2509.20146, 2025.

[89] Zixuan Shangguan, Yanjie Dong et al. Exploring and mitigating fawning hallucinations in large language models[J]. Neurocomputing, 2025: 132166.

[90] Huan Ma, Jiadong Pan et al. Semantic energy: Detecting LLM hallucination beyond entropy[J]. arXiv preprint arXiv:2508.14496, 2025.

[91] Sun X, Zhang D, Yang D, et al. Multi-turn context jailbreak attack on large language models from first principles[J]. arXiv preprint arXiv:2408.04686, 2024.

[92] Ohm M, Plate H, Sykosch A, et al. Backstabber’ s knife collection: A review of open source software supply chain attacks[C]//International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA). Springer, Cham, 2020.

[93] Zhang Z, Xiao G, Li Y, et al. Red alarm for pre-trained models: Universal vulnerability to neuron-level backdoor attacks[J]. Machine Intelligence Research, 2023

[94] Carlini N, Jagielski M, Choquette-Choo C A, et al. Poisoning web-scale training datasets is practical[C]//Proceedings of the IEEE Symposium on Security and Privacy (S&P). 2024: 407-425.

[95] Carlini N, Nasr M, Choquette-Choo C A, et al. Are aligned neural networks adversarially aligned?[J]. Advances in Neural Information Processing Systems (NeurIPS), 2023, 36: 61478-61500.

[96] Alber D A, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks[J]. Nature Medicine, 2025, 31(2): 618-626.

[97] Souly A, Rando J, Chapman E, et al. Poisoning attacks on LLMs require a near-constant number of poison samples[J]. arXiv preprint arXiv:2510.07192, 2025.

[98] Guo Z, Xu B, Zhu C, et al. MCP-AgentBench: Evaluating real-world language agent performance with MCP-mediated tools[J]. arXiv preprint arXiv:2509.09734, 2025.

[99] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models:

Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems (TOIS), 2025, 43(2): 1-55.

[100] Radosevich B, Halloran J. MCP safety audit: LLMs with the model context protocol allow major security exploits[J]. arXiv preprint arXiv:2504.03767, 2025.

[101] Hou X, Zhao Y, Wang S, et al. Model Context Protocol: Landscape, security threats, and future research directions[J]. arXiv preprint arXiv:2503.23278, 2025.

[102] Perez F, Ribeiro I. Ignore previous prompt: Attack techniques for language models[J]. arXiv preprint arXiv:2211.09527, 2022.

[103] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How does LLM safety training fail?[J]. Advances in Neural Information Processing Systems, 2023, 36: 80079-80110.

[104] Lv H, Wang X, Zhang Y, et al. CodeChameleon: Personalized encryption framework for jailbreaking large language models[J]. arXiv preprint arXiv:2402.16717, 2024.

[105] Ran D, Liu J, Gong Y, et al. JailbreakEval: An integrated toolkit for evaluating jailbreak attempts against large language models[C]//The Network and Distributed System Security. 2025.

[106] Zhou Y, Ni T, Lee W B, et al. A survey on backdoor threats in large language models (LLMs): Attacks, defenses, and evaluation methods[J]. Transactions on Artificial Intelligence, 2025: 3-3.

[107] Zhang B, Zhang Y, Ji J, et al. Safevla: Towards safety alignment of vision-language-action model via safe reinforcement learning[J]. arXiv preprint arXiv:2503.03480, 2025.

[108] Gong Y, Ran D, Liu J, et al. Figstep: Jailbreaking large vision-language models via typographic visual prompts[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(22): 23951-23959.

[109] Chen X, Tang S, Zhu R, et al. The janus interface: How fine-tuning in large language models amplifies the privacy risks[C]//Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security. 2024: 1285-1299.

[110] Chen S, Piet J, Sitawarin C, et al. StruQ: Defending against prompt injection with structured queries[C]//The 34th USENIX Security Symposium. 2025: 2383-2400.

[111] Yuan Y, Jiao W, Wang W, et al. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher[C]//The International Conference on Learning Representations. 2024.
[112] Liu X, Xu N, Chen M, et al. Autodan: Generating stealthy jailbreak prompts on aligned large language models[C]//The International Conference on Learning Representations. 2024.
[113] Huang Y, Sun L, Wang H, et al. Trustllm: Trustworthiness in large language models[C]//International Conference on Machine Learning. 2024.
[114] Zhang J, Liu D, Qian C, et al. Reef: Representation encoding fingerprints for large language models[C]//The International Conference on Learning Representations. 2024.
[115] Zeng B, Wang L, Hu Y, et al. Huref: Human-readable fingerprint for large language models[J]. Advances in Neural Information Processing Systems, 2024
[116] Tsai Y Y, Guo C, Yang J, et al. RoFL: Robust fingerprinting of language models[J]. arXiv preprint arXiv:2505.12682, 2025.
[117] Wu J, Peng W, Fu H, et al. ImF: Implicit fingerprint for large language models[J]. arXiv preprint arXiv:2503.21805, 2025.
[118] Zhang R. Matrix-driven instant review: Confident detection and reconstruction of LLM plagiarism on PC[J]. arXiv preprint arXiv:2508.06309, 2025.

国内标准

[1]2025 年 12 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 46800-2025《生成式人工智能技术应用社会影响 评估指南》
[2]2025 年 12 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 46799-2025《人工智能社会实验 评价指南》
[3]2025 年 10 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 46347-2025《人工智能 风险管理能力评估》
[4]2025 年 9 月；全国网络安全标准化技术委员会；TC260-004《政务大模型应用安全规范》
[5]2025 年 9 月；全国网络安全标准化技术委员会、国家计算机网络应急技术处理协调中心；《人工智能安全治理框架 2.0》

[6]2025 年 8 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 45958-2025《网络安全技术 人工智能计算平台安全框架》

[7]2025 年 6 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 45907-2025《人工智能 服务能力成熟度评估》

[8]2025 年 4 月；国家市场监督管理总局、国家标准化管理委员会；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》

[9]2025 年 4 月；国家市场监督管理总局、国家标准化管理委员会；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》

[10]2025 年 4 月；国家市场监督管理总局、国家标准化管理委员会；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》

[11]2025 年 4 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 45574-2025《数据安全技术 敏感个人信息处理安全要求》

[12]2025 年 2 月；国家市场监督管理总局、国家标准化管理委员会；GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》

[13]2025 年 2 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 45288.1-2025《人工智能 大模型 第 1 部分：通用要求》

[14]2025 年 2 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 45288.2-2025《人工智能 大模型 第 2 部分：评测指标与方法》

[15]2025 年 1 月；国家市场监督管理总局、国家标准化管理委员会；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]2024 年 11 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 45087-2024《人工智能 服务器系统性能测试方法》

[17]2024 年 2 月；全国网络安全标准化技术委员会；TC260-003《生成式人工智能服务安全基本要求》

[18]2023 年 5 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 42755-2023《人工智能 面向机器学习的数据标注规程》

[19]2022 年 10 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 41867-2022《信息技术 人工智能 术语》

[20]2022 年 10 月；国家市场监督管理总局、国家标准化管理委员会；GB/T 42018-2022

《信息技术 人工智能 平台计算资源规范》

国际标准

[1]2025 年 11 月；ISO/IEC；ISO/IEC TS 42119-2:2025 Artificial intelligence - Testing of AI - Part 2: Overview of testing AI systems

[2]2025 年 10 月；ISO/IEC；ISO/IEC 27701:2025 Information security, cybersecurity and privacy protection — Privacy information management systems — Requirements and guidance

[3]2025 年 9 月；ISO/IEC；ISO/IEC TS 6254:2025 Information technology - Artificial intelligence - Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems

[4]2025 年 7 月；ISO/IEC；ISO/IEC 42006:2025 Information technology - Artificial intelligence - Requirements for bodies providing audit and certification of artificial intelligence management systems

[5]2025 年 7 月；WDTA；Single AI Agent Runtime Security Testing Standards

[6]2025 年 6 月；ISO/IEC；ISO/IEC 25389:2025 Information technology - The safe framework

[7]2025 年 5 月；ISO/IEC；ISO/IEC 42005:2025 Information technology - Artificial intelligence (AI) - AI system impact assessment

[8]2025 年 2 月；ISO/IEC；ISO/IEC 5259-5:2025 Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 5: Data quality governance framework

[9]2024 年 4 月；WDTA；Large Language Model Security Testing Method

[10]2024 年 4 月；WDTA；Generative AI Application Security Testing and Validation Standard

[11]2024 年 1 月；ISO/IEC；ISO/IEC 5339:2024 Information technology - Artificial intelligence - Guidance for AI applications

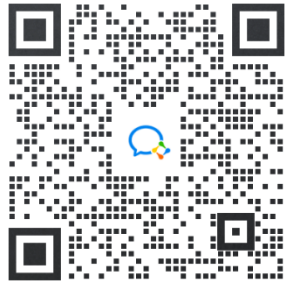
[12]2024 年 1 月；ISO/IEC；ISO/IEC TR 5469:2024 Artificial intelligence - Functional safety and AI systems

[13]2024 年 1 月；ISO/IEC；ISO/IEC TS 25058:2024 Systems and software engineering -

免责申明：

1. 本资料来源于网络公开渠道，版权归属版权方；
2. 本资料仅限会员学习使用，如他用请联系版权方；
3. 会员费用作为信息收集整理及运营之必须费用；
4. 如侵犯您的合法权益，请联系客服微信将及时删除。

行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息

微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享10000+份行业研究报告、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等全领域；是全网分享数量最多、质量最高、更新最快的知识社群。

加入后无限制搜索下载

Systems and software Quality Requirements and Evaluation (SQuaRE) - Guidance for quality evaluation of artificial intelligence (AI) systems
[14]2023 年 12 月； ISO/IEC； ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system
[15]2023 年 12 月； ISO/IEC； ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes
[16]2023 年 5 月； ISO/IEC； ISO/IEC TR 27563:2023 Security and privacy in artificial intelligence use cases - Best practices
[17]2023 年 2 月； ISO/IEC； ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management
[18]2023 年 1 月； NIST； Artificial Intelligence Risk Management Framework (AI RMF 1.0)
[19]2022 年 12 月； DIN & DKE； GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE 2ND EDITION
[20]2022 年 10 月; ISO/IEC; ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements
[21]2022 年 7 月； ISO/IEC； ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
[22]2022 年 2 月； OECD； FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS