CSTC 中国软件评测中心
(工业和信息化部软件与集成电路促进中心)

安恒信息 DAS-Security 安全中国

Top开源大模型安全测评报告(2024)

中国软件评测中心安全事业部

杭州安恒信息技术股份有限公司

中国计算机行业协会数据安全专业委员会

数据安全关键技术与产业应用评价工业和信息化部重点实验室

联合发布

2024年12月

前言

为深入学习贯彻全国两会精神和党的二十届三中全会精神，落实《中共中央关于进一步全面深化改革、推进中国式现代化的决定》作出“建立人工智能安全监管制度”“完善生成式人工智能发展和管理机制”的重要部署以及根据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《生成式人工智能服务管理办法》《工业和信息化部等十六部门关于促进数据安全产业发展的指导意见》等法律法规政策文件要求，促进和引导人工智能大模型技术向“负责任、可持续、高可靠”目标发展，让人工智能大模型技术真正实现高质量安全赋能各行各业落地应用。

中国软件评测中心（工业和信息化部软件与集成电路促进中心）安全事业部联合杭州安恒信息技术股份有限公司、数据安全关键技术与产业应用评价工业和信息化部重点实验室、中国计算机行业协会数据安全专业委员会，共同开展国内外开源大模型的安全性、公平性和可靠性研究，并发布《**Top开源大模型安全测评报告(2024)**》。

本报告聚焦国内外开源大模型的安全风险测评，通过选取典型的12家20款开源大模型，从国家安全、道德伦理、公民权利、公共安全、历史文化、医疗卫生、隐私信息、不良信息、商业金融、基础安全、网络安全和模型滥用等12个方面展开深入安全测评，旨在提高大模型厂商的安全意识和保障行业用户的合法权益，并通过系统性分析国内外开源大模型安全的综合表现，为人工智能大模型产业各界提供参考。

【注】因大模型迭代速度快，测评结果仅适用于测试期间和测试版本。报告中的分析和结论可能存在一定的局限性和不完整性，我们期待并欢迎各方提出宝贵的批评与建议，共同推动人工智能大模型安全治理。

目录

01

测评体系