京东

Agent开发者论坛

2025京东全球科技探索者大会

AI时代下安全新范式：JoySafety + 安全Agent

AI时代安全风险和挑战

JoySafety：AI的“守护者”，化解其原生风险

新的安全风险

- 提示注入2.0：从“聊天越狱”到“Agent劫持
- AI投毒：从“数据”到“恶意代码/MCP工具”

安全Agent：传统安全的“创新者”，重塑防御体系

给传统安全带来的挑战

- 攻击手段更智能：攻击的门槛降低、规模变大，更持续
- 风险面扩大：新型数据泄漏、内容安全
- 响应时效：攻防节奏从“回合战”转向“实时战”

JoySafety：全链守护AI

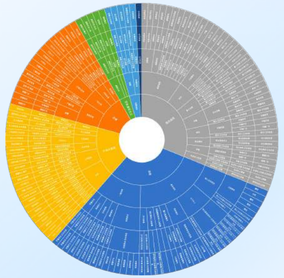
全链守护AI


内容安全风险
业务安全风险
信息安全风险
模型训练层
数据投毒零容忍
模型评测层
31类风险基线
模型运营层
实时风险识别
第一道防线
第二道防线
第三道防线
第四道防线
训练数据安全
大模型安全评测
Prompt实时检测
生成内容实时检测

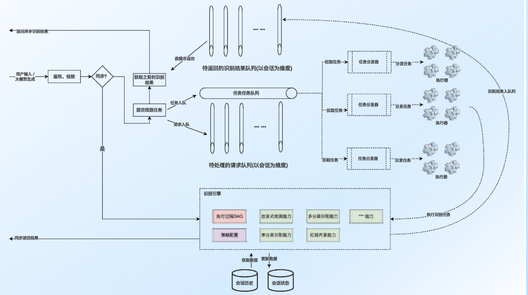
全自动、一站式合规

多种响应方式

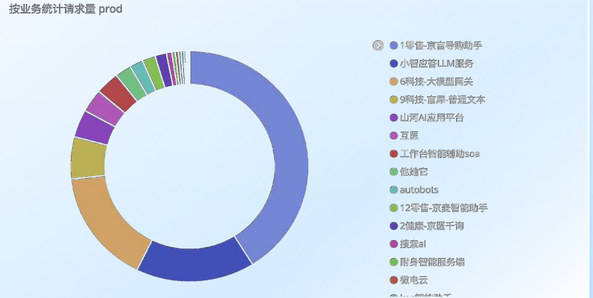
31大类+200+子类安全风险识别



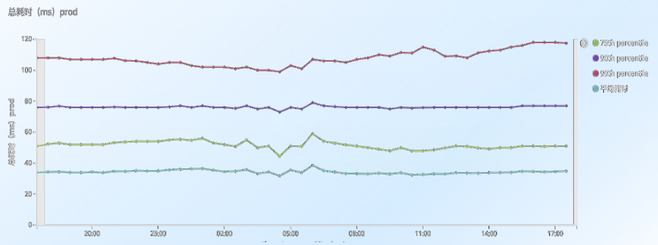
“流式输出检测 + 撤回”机制



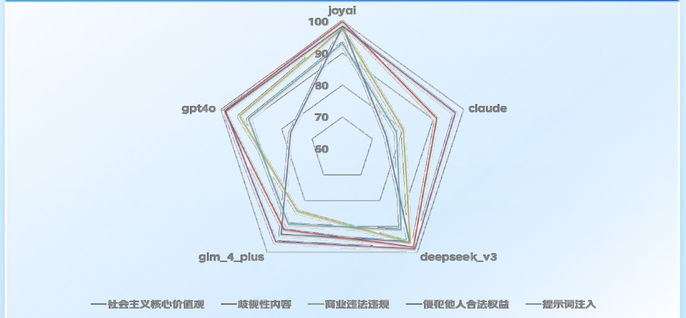
100+应用/亿级请求



毫秒级响应

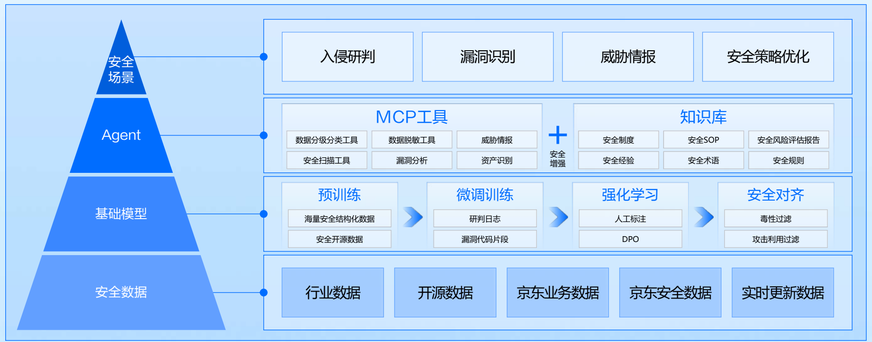


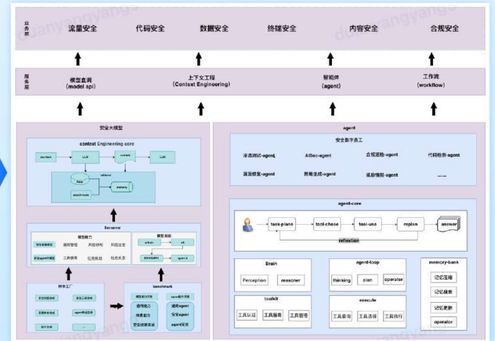
降低攻击95%以上



安全Agent：安全数字员工

JSL安全大模型2.0： Model as Agent





模型	数学 GSM8K	代码 EvalPlus	指令遵循 IFEval	安全通识	代码安全	内容安全	终端安全	威胁情报	流量安全
JSL-V1	70.13	69.50	72.09	56.08	46.56	74.24	82.35	51.50	53.33
JSL-V2	94.72	86.63	**83.69**	**77.68**	**72.03**	**80.71**	**91.43**	**86.67**	**75.47**
SecGPT-14B	81.80	76.80	64.33	64.10	42.71	60.18	59.46	61.25	48.81
Qwen3-32B-Think	**94.84**	89.02	83.18	72.26	51.20	70.29	60.46	65.25	53.73
DeepSeek-R1	89.00	**93.30**	83.30	79.71	–	–	–	82.38	61.95

注：加粗表示该列最优值；“–”因为安全原因暂时未评测，↑越高越好。

JSL-CodeSafeter： 代码漏洞检测与修复Agent

VS

漏洞检出阶段滞后 代码提测甚至上线阶段才能检出漏洞	风险发现阶段	**安全左移，入库即安全** 在代码编写阶段实时拦截漏洞，杜绝“带病入库”
平均修复周期长，预计为7天 安全团队扫描 → 人工分析→ 提单给研发 → 研发排期修复	平均修复周期	**修复周期缩短至30min** 漏洞实时发现 → 自动修复方案生成→ 研发10分钟内确认
学习成本高 安全专家需解释漏洞原理、提供修复方案；研发人员因安全知识不足常产生抵触	漏洞修复方式	**“傻瓜式”修复指引** 直接列出修复前后代码，研发直接点击接受即可

代码编写阶段进行代码安全扫描，实现“人防”到“技防”的跨越

CWE TOP25+	90%	1W+	30%+
漏洞覆盖	识别准召	覆盖用户	修复采纳

JSL-PenTester： 渗透测试Agent

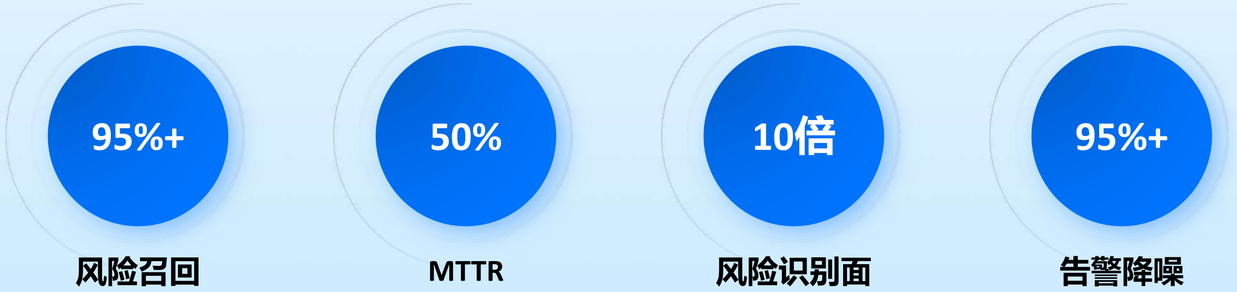
AI驱动	VS	人工驱动
◎ 由AI算法驱动，能够自主执行测试流程 ◎ 效率指数级提升： 7x24小时不间断	执行主体及效率	高度依赖安全专家（白帽黑客）的个人经验、知识和直觉 ◎ 效率瓶颈明显 ◎
◎ 具备“全知”潜力和能力 ◎ 永不遗忘且持续进化	知识广度与深度	测试深度受限于工程师的知识盲区 ◎ 知识更新依赖人工学习 ◎
◎ 可实现“持续性”安全验证 ◎ 追求全覆盖	覆盖范围与持续性	对一个“时间点”的定点快照测试 ◎ 样本覆盖有限 ◎

“多Agent 协同--多工具集成—自我反思迭代”实现全流程智能化

Owasp Top 10	从「天」降低至「小时」	-70%	+30%
漏洞覆盖	效率	专家接入率	漏洞发现率

JSL-AlertTriager： 入侵研判Agent

维度	传统SOC告警研判	AI SOC智能研判Agent
核心驱动	人工流程与静态规则	AI模型与自动化
检测能力	仅已知威胁（基于规则）	已知+未知威胁（基于异常）
响应速度	慢（分钟/小时级，人工）	快（毫秒/秒级，自动）
人员角色	告警操作工，疲劳度高	战略调查员，价值提升
误报率	极高（常 >90%）	低（可降至10%以下）



JSL-ThreatIntelligerː 威胁情报Agent

亿级/天

线索数量

3min

MTTR

98%

情报新鲜度

80%+

情报有效率

AI时代下安全展望

JoySafety

作为“AI守护者”为大模型的安全运行和合规使用提供保障

JSLSafeter

传统安全的“创新者”，重塑信息安全防御体系

JLBoost

大模型的“助力器”，在大模型训练数据获取、幻觉方面提供支撑，提升大模型输出可靠性和可信度。

携手共创可信AI

开源共建：携手共创可信AI

共享

共创

共建

第一步：快速接入

访问GitHub、 huggingface 获取JoySafety代码及模型，分钟级快速接入。
https://github.com/jd-opensource/JoySafety/tree/dev
https://huggingface.co/jdopensource/JoySafety

第二步：共创共建

加入官方微信，提交场景需求及建议，共创共建可信AI。
所有开源组件遵循Apache 2.0协议

第三步：持续优化

诚邀开发者、企业、科研机构一起贡献代码、数据与场景，持续打磨面向未来的AI安全范式

谢谢