象信AI安全护栏

基于大模型的开源AI安全护栏

🎯 企业级AI安全防护解决方案

支持完全私有化部署 • 开源透明 • 基于大模型技术

象信AI公司介绍

🎯 公司定位

开源的AI安全护栏，企业级防护解决方案提供商

🚀 核心使命

让象信AI安全护栏成为AI安全的标配

我们是做什么的

- **安全网关模式和API检测模型** – 多种接入方式灵活部署
- **防护提示词攻击检测和中文内容安全** – 专业安全防护
- **基于LLM大模型具备上下文感知能力** – 智能语义分析
- **支持完全私有化部署** – 数据安全可控

🌟 开源贡献

- 开源了AI安全护栏的基础安全能力大模型
- 开源了企业级AI安全管理平台
- 活跃的开源社区支持

创始人介绍

王磊 – 创始人兼CEO

- 🎓 天津大学 计算机科学与技术
- 💼 百度工作十年
- ⚖️ 国家法律职业资格
- 🏛️ 工信部重点实验室特聘技术专家

🚀 职业历程

2014年 – 作为安全专家接受中央电视台《晚间新闻》采访，向全国手机用户解析超级病毒"短信蠕虫"的危害及传播路径

2016年 – 在吴恩达的指导下，AI驱动的移动端恶意代码检测技术论文登上国际安全会议Blackhat

2021年 – AI驱动的数据隐私合规项目入围百度百万美金最高奖，同年取得国家法律职业资格

2023年3月 – 离开百度，创立北京象信智能科技有限公司，致力于AI安全技术创新

2024年1月 – 受聘为工业和信息化部重点实验室特聘技术专家

🎯 技术与法律并重

王磊拥有深厚的AI技术背景和法律专业知识，这种跨领域的专业素养使他能够从技术实现和合规监管两个维度深入理解AI安全的本质需求，为象信AI的产品研发和战略发展奠定了坚实基础。

什么是AI安全护栏

定义

又称大模型安全护栏、大模型防火墙，用于限制大模型应用输入输出内容或行为，防止重要数据泄露、提示词注入攻击、生成违法和不良信息

核心功能

🔍 输入安全检测

识别和过滤恶意输入，防范注入攻击

📋 输出内容审核

确保模型输出符合安全规范

🛡️ 提示词攻击防护

防范注入、越狱等各类攻击

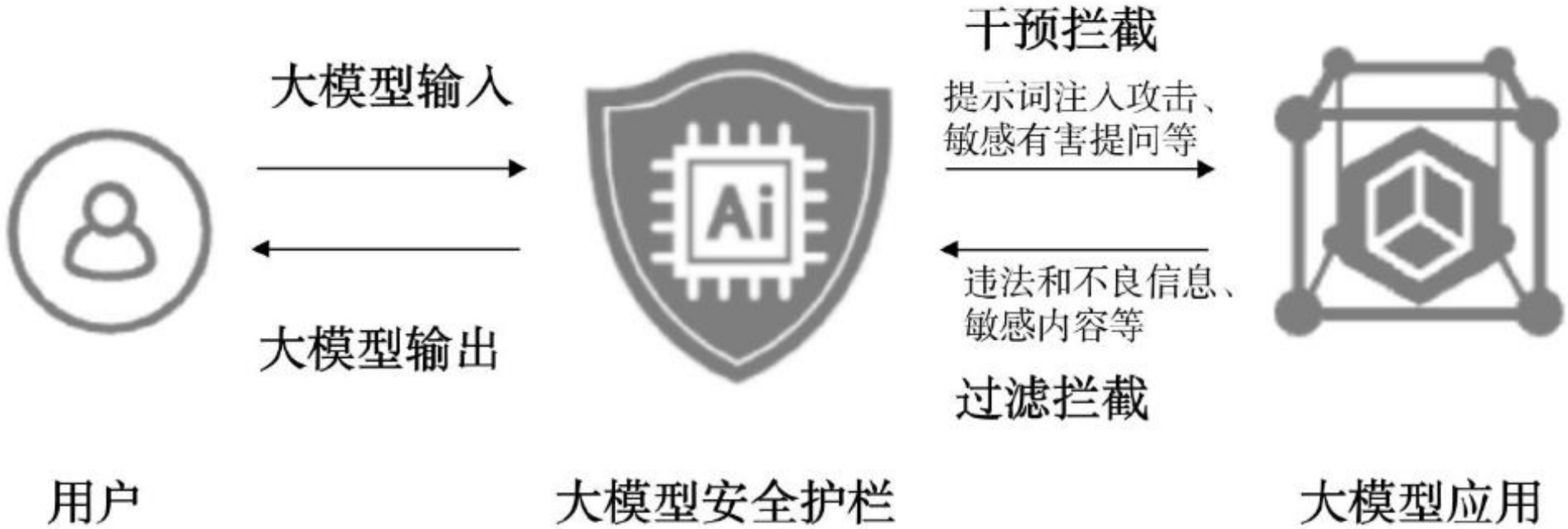
📊 行为审计

完整记录和分析用户行为

技术特点

- 基于LLM大模型的上下文感知能力
- 支持多种部署模式（智能体编排、网关代理、直接串联）
- 实时安全检测和响应

大模型安全护栏工作原理



干预拦截

- 提示词注入攻击
- 敏感有害提问等

过滤拦截

- 违法和不良信息
- 敏感内容等

为什么需要大模型安全护栏

📜 法规合规要求

- 国家网信办《生成式人工智能服务管理暂行办法》
- AI应用对公众提供服务（大模型备案登记备案）的安全要求
- 国务院"人工智能+"行动安全发展要求

⚠️ 安全风险挑战

- 提示词攻击（注入、越狱）
- 内容安全（违法违规信息）
- 数据泄露风险
- 恶意利用威胁

100%

恶意攻击拦截率

0

敏感信息泄露

7x24

全天候防护

企业部署AI应用必须考虑安全合规，象信AI安全护栏是您的最佳选择

象信AI安全护栏的优势

🧠 技术优势

- 基于大模型的语义理解能力
- 开源架构，透明可审计
- 中文内容安全专项优化
- 上下文感知和连贯性分析

🚀 部署优势

- 多种模式：网关、串联、一体机
- 完全私有化部署
- 华为昇腾NPU信创支持
- 智能体编排平台集成

🔒 安全能力

- 覆盖输入输出全链路防护
- 毫秒级实时检测响应
- OWASP Top 10 LLM Applications 2025
- 符合国标安全要求

💼 商业优势

- 开源免费降低成本
- 灵活的商业授权模式
- 完善的技术支持服务
- 活跃的社区生态

版本策略

版本	SaaS版	社区版	企业版	行业版
适用对象	个人开发者 中小微团队	中小型团队 和企业	中大型企业	特定监管要求 行业客户
价格	免费+增值	免费	联系销售	联系销售
部署方式	官网在线服务	开源软件自部署	厂家部署	厂家部署
技术支持	原厂在线支持	社区技术支持	5x8 基础级支持	7x24 增强级支持

🎯 **针对不同规模和需求的企业，提供最适合的安全防护方案**

核心功能对比 – 安全防护能力

🧠 安全能力大模型

- ✅ **标准安全能力**：提示词攻击防护+内容安全+个人信息安全（全版本支持）
- ✅ **私有数据精调**：基于客户私有数据精调安全能力（企业版+行业版）
- ✅ **行业特定安全**：符合行业特定要求的安全能力（行业版专享）

🛡️ 提示词攻击防护（根据《OWASP Top 10 for LLM Applications 2025》）

注入攻击防护

直接注入、间接注入、递归注入、系统提示词泄漏等

越狱攻击防护

角色扮演、输入混淆、上下文操纵、假定场景等

资源消耗攻击

随机字符、重复令牌、过度推理等

恶意操作防护

SQL注入、命令注入、XSS、SSRF、恶意代码等

核心功能对比 – 内容安全与审计

📜 内容安全（根据《GB/T45654—2025 生成式人工智能服务安全基本要求》）

A.1 价值观安全

违反社会主义核心价值观内容识别

A.2 歧视性内容

各类歧视性和不当内容识别

A.3 商业合规

商业违法违规内容检测

A.4 权益保护

侵犯他人合法权益内容识别

🔒 数据安全 & 📄 合规审计

- ✅ **个人信息安全**：根据《GB/T35273—2020 个人信息安全规范》（全版本支持）
- ✅ **商业秘密**：全面商业秘密保护（行业版专享）
- ✅ **涉密信息**：全面涉密信息保护（行业版专享）
- ✅ **行业重要数据**：根据行业相关法律法规（行业版专享）
- ✅ **日志留存**：完整的日志留存和审计能力（全版本支持）
- ✅ **一键审计报告**：符合多个监管部门要求（行业版专享）

部署模式与技术支持

🏢 部署模式（全版本支持）

智能体编排

支持Dify、Coze等主流平台无缝集成

网关代理

透明接入现有系统，无需修改应用代码

直接串联

与应用紧密集成，API直接调用

软硬一体机

开箱即用的企业级解决方案（企业版+）

🇨🇳 信创大模型支持

- Nvidia GPU（全版本支持）
- 华为昇腾NPU（行业版专享）

🤝 服务规格

- 基础级：5x8支持，4小时响应
- 增强级：7x24支持，1小时响应

详细功能对比

功能特性	SaaS版	社区版	企业版	行业版
🧠 安全能力大模型				
标准安全能力（提示词攻击防护+内容安全+个人信息安全）	✅	✅	✅	✅
基于客户私有数据精调安全能力	❌	❌	✅	✅
符合客户所属行业特定要求的安全能力	❌	❌	❌	✅
🛡️ 提示词攻击防护（根据《OWASP Top 10 for LLM Applications 2025》）				
提示词注入攻击（直接注入、间接注入、递归注入、系统提示词泄漏等）	✅	✅	✅	✅
越狱攻击（角色扮演、输入混淆、上下文操纵、假定场景、反向诱导、对抗性攻击等）	✅	✅	✅	✅
资源消耗攻击（随机字符、重复令牌、过度推理等）	✅	✅	✅	✅
恶意操作（SQL注入、MCP工具投毒、命令代码注入、XSS、SSRF、恶意代码、路径遍历等）	✅	✅	✅	✅

详细功能对比

功能特性	SaaS版	社区版	企业版	行业版
🔒 内容安全（根据《GB/T45654—2025 生成式人工智能服务安全基本要求》）				
A.1 包含违反社会主义核心价值观的内容	✅	✅	✅	✅
A.2 包含歧视性内容	✅	✅	✅	✅
A.3 商业违法违规	✅	✅	✅	✅
A.4 侵犯他人合法权益	✅	✅	✅	✅
A.5 无法满足特定服务类型的安全需求	❌	❌	❌	✅
🛡️ 数据安全				
个人信息安全（根据《GB/T35273—2020 个人信息安全规范》）	✅	✅	✅	✅
商业秘密（根据国家商业秘密相关法规及企业内部管理制度）	❌	❌	❌	✅
涉密信息（根据国家保密法律法规及相关标准要求）	❌	❌	❌	✅
行业重要数据（根据行业相关法律法规及相关标准要求）	❌	❌	❌	✅

详细功能对比

功能特性	SaaS版	社区版	企业版	行业版
🔍 大模型输入风险识别管控				
支持语义级分析能力，可自动识别分类违法和不良信息	✅	✅	✅	✅
自定义关键词过滤规则	✅	✅	✅	✅
支持上下文关联分析，可对超长会话历史进行连贯性分析	✅	✅	✅	✅
自动识别/拦截个人信息等敏感内容	✅	✅	✅	✅
基于用户角色识别/拦截越权提问信息	❌	❌	❌	✅
基于业务场景识别/拦截超范围提问信息	❌	❌	❌	✅

详细功能对比

功能特性	SaaS版	社区版	企业版	行业版
📤 大模型输出风险识别管控能力				
识别过滤大模型输出的违法和不良信息	✅	✅	✅	✅
拒答答案库管理（可配置、可扩展、可更新）	✅	✅	✅	✅
代答知识库管理（可配置、可扩展、可更新）	✅	✅	✅	✅
配置脱敏规则，对大模型生成的敏感内容进行脱敏后输出	❌	❌	❌	✅
基于业务场景识别/过滤大模型输出超范围内容	❌	❌	❌	✅
📋 合规审计				
具备日志留存和审计能力	✅	✅	✅	✅
一键出审，生成符合多个监管部门要求的审计报告	❌	❌	❌	✅

详细功能对比

功能特性	SaaS版	社区版	企业版	行业版
🎯 支持的模态				
文本识别	✅	✅	✅	✅
图像识别	❌	❌	❌	✅
音频识别	❌	❌	❌	✅
视频识别	❌	❌	❌	✅
文件识别	❌	❌	❌	✅
🖥️ 部署模式				
智能体编排（Dify, Coze等）	✅	✅	✅	✅
网关代理	✅	✅	✅	✅
直接串联	✅	✅	✅	✅
软硬一体机交付（Nvidia或信创服务器，不含硬件费用）	❌	❌	✅	✅

详细功能对比

功能特性	SaaS版	社区版	企业版	行业版
🇨🇳 信创大模型				
Nvidia GPU	✅	✅	✅	✅
华为昇腾NPU	❌	❌	❌	✅
🤝 服务规格				
开源社区支持	✅	✅	✅	✅
原厂级企业级支持服务	线上支持	❌	基础级	增强级

成功案例

🚗 网约车出行平台

挑战：AI客服交互合规性与安全性
方案：企业版安全护栏
效果：开源免费降低成本，有效过滤不当内容

🏦 金融行业

挑战：客服机器人安全合规
方案：金融行业版AI安全护栏
效果：100%拦截恶意攻击，零敏感信息泄露

🏛️ 政务服务

挑战：政府服务热线内容管控
方案：政务行业版+华为昇腾NPU
效果：完全私有化，满足政务安全规范

🎓 教育科技

挑战：在线教育AI内容审核
方案：教育行业版安全护栏
效果：满足教育部门监管，保护学生健康

🏆 各行各业都需安全护栏来保护AI应用

开源生态与技术实力

🌟 开源模型

Hugging Face & ModelScope

xiangxinai/Xiangxin-Guardrails-Text

💻 开源平台

GitHub

xiangxinai/xiangxin-guardrails

企业级AI安全管理平台完全开源

技术架构优势

- **模型驱动**：基于自研大模型，具备强大的语义理解能力
- **上下文感知**：支持超长会话历史的连贯性分析
- **实时检测**：毫秒级响应，不影响用户体验
- **多模态支持**：文本、图像、音频、视频全覆盖（行业版）

30000+

开源代码量

100000+

开源模型累计下载量

10+

行业解决方案

竞争优势分析

🆚 象信AI vs 传统方案

象信AI优势：

- 基于大模型的语义理解
- 开源透明，成本更低
- 中文内容安全专项优化
- 完全私有化部署

📊 传统方案局限

传统方案问题：

- 基于规则，误报率高
- 闭源黑盒，不可定制
- 英文为主，中文效果差
- 依赖云服务，数据不可控

核心差异化优势

🧠 技术领先

国内首个开源的基于大模型的AI安全护栏

🎯 符合国家标准

严格依照国家标准校准安全能力

🔓 开源透明

完全开源，客户可审计、可定制

🛡️ 信创支持

支持华为昇腾等国产硬件

实施路线图

1 → 2 → 3 → 4

1 大模型应用安全测评

- 业务场景分析
- 执行安全测评

2 安全评估报告

- 生成整改建议
- 生成安全评估报告

3 部署实施

- 环境搭建
- 系统安装

4 上线运维

- 系统上线
- 监控告警

实施周期

版本	SaaS版	社区版	企业版	行业版
部署周期	即开即用	1–2天	3–5天	1–2周
安全能力大模型	标准安全能力	标准安全能力	私有数据精调安全能力	行业特定安全能力要求+产品功能要求
培训服务	在线文档	社区支持	现场培训	专业培训

联系我们

🌐 官方信息

官网：https://xiangxinai.cn

GitHub：https://github.com/xiangxinai/xiangxin-guardrails

Hugging Face：xiangxinai/Xiangxin-Guardrails-Text

📞 商务联系

联系人： 王先生

邮箱： wanglei@xiangxinai.cn

电话： 18618148202

🤝 合作方式

社区版： 直接下载使用

企业版： 联系商务洽谈

行业版： 联系商务洽谈

象信AI安全护栏

成为AI时代的安全标准配置