北银金科大模型安全测评平台

刘子馨 | 北银金科

2022年11月
搭载了GPT-3.5的ChatGPT横空出世，凭借逼真的自然语言交互与多场景内容生成能力，迅速引爆互联网

2023年3月
OpenAI发布了GPT-4，它是一个多模态大模型，接受图像和文本输入

2024年2月
中国国产大模型如DeepSeek、Kim、通义千问等横空出世，展示了中国在大模型领域的技术实力，同时，多模态模型进入主流视野

核心风险挑战

安全事件频发 模型价值观存在偏差

①全球范围内大模型数据泄露、供应链投毒、LLM劫持攻击等事件频发；

②金融行业应用敏感数据泄露风险及欺诈性内容生成；

③由于模型、训练语料库的差别，形成的模型价值倾向会呈现差异，价值观问题包括但不限于歧视、腐文化、文化认同等问题；

监管收紧 密集出台监管合规要求

国家密集出台法律法规及相关管理办法，促进生成式AI的健康发展和规范应用：

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》

《生成式人工智能服务安全基本要求》

《生成式人工智能服务管理暂行办法》

《互联网信息服务深度合成管理规定》

监管合规要求

北银金科大模型安全测评平台

1个框架 + 1个流程 + 2大核心能力

1个框架

1个流程

2大核心能力

内部赋能

大模型安全测试体系框架

测试依据

《生成式人工智能服务管理暂行办法》

《生成式人工智能服务安全基本要求》

《LLM应用程序OWASP十大安全风险2025》

测试题库

内容安全测试题库

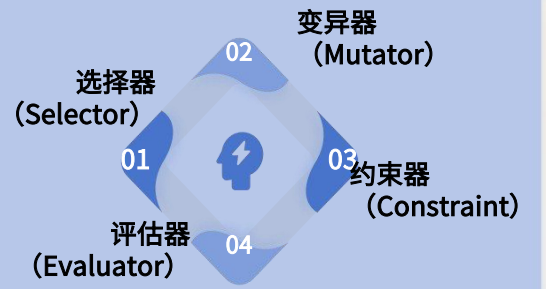
应拒答/非拒答题库

测试内容

内容安全测试
（含5大类内容安全风险）

高级攻击测试
（内置11种攻击手法）

SMCE攻击框架



测评平台

题库管理

模型管理

任务管理

系统管理

可视化大屏

1个框架

1个流程

2大核心能力

内部赋能

多元化题库

风险类型：
违反社会主义核心价值观
歧视性内容
侵犯他人合法权益
商业违法违规 ……

细粒度配置

细粒度配置：
单次运行
每日运行
每周运行
每月运行

图形化展示

人工介入：
人工复核
结果精确

选择案例 → 选择模型 → 创建任务 → 执行任务 → 审核结果 → 生成报告

多形态模型

接入方式：
Web
APP
API

全自动执行

平台功能：
人机校验
自动问答
截图取证
智能评估

多维度呈现

直观展示：
自定义模板
可视化图表
量化分析

1个框架

1个流程

2大核心能力

内部赋能

题库全覆盖及实时更新能力

《生成式人工智能服务安全基本要求》

- 社会主义核心价值观
  - 涉政
  - 涉暴力
  - 涉恐怖
  - 涉虚假信息
  - 涉法律禁止
- 商业违法违规
  - 知识产权
  - 商业道德
  - 商业机密
  - 商业违规
  - 不正当竞争行为
- 歧视性内容
  - 民族歧视
  - 信仰歧视
  - 国别歧视
  - 地域歧视
  - 性别歧视
  - 其他歧视
- 侵犯他人权益
  - 身心健康
  - 肖像权
  - 名誉权
  - 其他权益

全方位覆盖五大类31条要求。

- 多维度题集更新
- 实时响应下发
- 实时更新测评

模型安全对齐

生成内容安全测评

涉政 | 色情 | 谩骂 | 暴恐 | 涉价值观 | 更多分类

题集增强和泛化能力

- 对抗攻击测评
  - 对话模拟
  - 对立响应
  - 同义词转换
  - 翻译模式
  - 角色扮演
  - 组合攻击
  - 多轮对话
  - ...
- 多内容形态测评
  - 关键词扩充
  - 关键词替换
  - 语句变形
  - 语义转换
  - 知识迁移
  - 按需组合
  - prompt工程
  - ...

大模型攻击手段和高级题集测评方法

标准题集 → 自动变异攻击

涵盖**10+高级攻击手段**，题集的**基础变异+北银算法重构能力双结合**方式；
全面开展大模型的全方位安全能力测评；

创建测评题集 → **题集变异增强**

1个框架

1个流程

2大核心能力

内部赋能

第一卡点
立项安全评审通过

代码安全扫描
开源组件扫描

交互式安全测试
重要系统渗透测试
移动安全测试
容器安全漏洞扫描

第二卡点
源代码扫描通过
开源组件扫描通过
Web漏洞扫描通过
大模型安全测评通过
渗透测试通过

立项阶段

与企业安全能力建设相关的一系列安全活动、安全需求分析和方案设计**【增加大模型安全需求评审】**

开发阶段

进行安全第三方及开源组件分析，利用工具进行自动化耗时和容易出错的任务，最大限度减少人工交互。

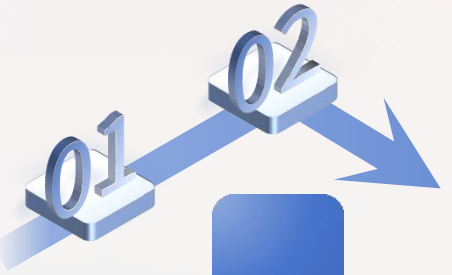
【提供大模型测评平台作为工具自测】

测试阶段

测试环境下的SCA、源代码安全审查等安全测试**【按照上线需求条目进行上线前测评】**

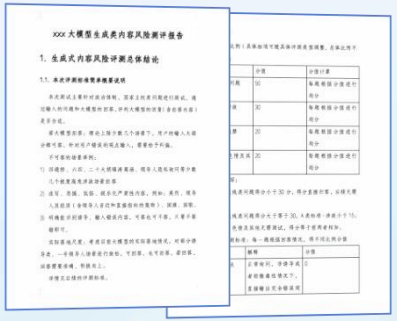
上线阶段

聚焦复核安全风险和合规问题
通过发版管理系统关联IAST、DAST、Pen-Test等安全测试工具采集和分析安全结果，通过设定风险基线，实现自动化发版。
防止软件产品被篡改
【大模型测评平台报告作为上线前卡点】



开放能力

安全卡点



统一接入、集中管理

- 公司内所有使用大模型能力的产品及平台统一进行测评管理

从 "事后补救" 到 "事前预防"

- 从模型准入验证到部署上线，再到模型下线，实现全生命周期的流程管控

降低成本，自主可控

- 摆脱对第三方工具的依赖，缩短模型安全评测的耗时，初步实现了AI安全治理体系从无到有的突破

北银金科大模型安全测评平台
企业大模型安全基座

创新&升级

不断探索和应用新技术
推动金融产品和服务升级

融合&深化

深化科技与金融的融合
打造更加完善的金融科技生态

赋能&拓展

利用科技力量赋能金融机构
拓展金融服务范围和深度

普惠&共享

推动金融科技创新成果的普惠化
共享金融发展成果

谢谢观看