布兰矩阵 AI 安全防护大模型产品/解决方案

白皮书

文档版本：V1.1

发布日期：2026 年 5 月

【服务声明】

本文档用于介绍 AI 安全防护大模型防护模型的技术定位、方法依据和适用场景。除双方商业合同另有约定外，本文档不构成对具体产品功能、部署规格、性能指标、合规结果或服务等级的承诺。

1 【修订记录】

编号	修改日期	修改人	更新内容
01	2026 年 5 月	布兰矩阵 AI 安全实验室	第一次发布。
02	2026 年 5 月	布兰矩阵 AI 安全实验室	扩充数据集构建与 SFT 训练方法描述。

2 目 录

3 引言

3.1 编写目的

本文档为 布兰矩阵智能科技 (上海) 有限公司 布兰矩阵 AI 安全实验室 提供的《AI 安全防护大模型 AI 对话防护模型》的产品白皮书。

3.2 文档使用说明

本文档用于介绍 AI 安全防护大模型 AI 对话防护模型，包括产品简介、产品定位、产品优势、方案架构、功能特性、应用场景和部署方案等方面的描述，方便客户相关人员了解该项产品与服务。

本文档中的模型规模、训练数据和评测数据采用概括性表述。具体版本能力、服务标准、交付内容和验收指标以双方合同、技术附件和实测报告为准。

3.3 术语和定义

LLM：Large Language Model，大语言模型

Prompt：用户向大语言模型输入的请求或提示

Response：AI assistant 针对 Prompt 给出的回复

AI 防护模型：面向大模型输入和输出的安全判定模型

AI 安全防护大模型：布兰矩阵基于自有安全数据训练的、采用 CoT（思维链）的 AI 对话防护模型

CoT：Chain-of-Thought，思维链，模型在给出最终结论前显式输出的中间推理步骤，是 AI 安全防护大模型可解释判定的核心机制

CoT 监督微调：使用安全标签与思维链推理过程共同训练模型，使模型在判定前形成可复核的 CoT 推理链

困难样本偏好优化：针对边界样本和易混淆样本进一步优化模型的判定一致性

用户输入有害性审计：判断用户请求是否包含违法违规、攻击、隐私、伦理等风险

模型回复有害性审计：结合用户请求和模型回复，判断回复内容是否存在安全风险

拒答行为审计：判断模型回复是合理拒答还是遵从请求，用于评估对齐和过度拒答问题

4 解决方案概述

4.1 产品简介

随着大语言模型在智能客服、办公助手、知识库问答、内容生成、代码辅助和 Agent 自动化等场景中落地，企业应用面临新的安全和合规挑战。用户输入可能包含越权指令、提示注入、越狱诱导、隐私探针、违法违规请求或高风险咨询；模型回复也可能在被诱导后输出不当内容、泄露敏感信息或给出不符合业务边界的建议。

传统关键词、正则、黑白名单等方式难以充分覆盖大模型对话中的语义变体和对抗表达。AI 安全防护大模型通过自有安全数据训练对话防护模型，在一次审计中同时关注用户输入、模型回复和拒答行为，为大模型应用提供输入侧拦截、输出侧审计和离线安全评测等基础能力。

AI 安全防护大模型已基于 Qwen3.5-9B-Base、Qwen3.5-4B-Base、Qwen3.5-0.8B-Base 训练三个版本，便于在不同业务场景中按能力、成本和资源约束进行选型。

4.2 产品定位

AI 安全防护大模型是面向大语言模型应用场景的 AI 对话防护模型，聚焦于用户输入风险检测、模型回复风险检测和拒答行为审计。AI 安全防护大模型可作为生成模型上游的输入侧防护组件，也可作为模型输出后的审计组件，还可用于研发测试阶段的安全评测。

产品交付可根据客户数据安全要求、模型调用链路和运维条件选择私有化、离线评测或接口集成等方式。具体部署形态、接口协议、数据留存和服务等级以项目方案和合同约定为准。

4.3 产品优势

三维审计能力：同时覆盖用户请求有害性、模型回复有害性和拒答行为，避免只看输入或只看输出导致的审计盲区。

基于 CoT 的可解释输出：模型在给出结论前先输出 CoT 思维链，便于业务侧复核、误判定位和后续样本沉

淀。

自有数据训练：训练数据来自布兰矩阵自有安全数据，可围绕中文业务场景、行业表达和真实对话形态持续迭代。

多规格模型选型：提供 9B、4B、0.5B 三档模型版本，便于在效果、成本、部署资源和响应要求之间平衡。

可扩展审计基线：风险类型、处置策略和复核流程可按客户业务线进行配置，支持后续接入更多行业场景样本。

4.4 建设目标

提供面向大模型对话场景的输入侧与输出侧安全审计能力

提供可解释、可复核、可沉淀的模型判定结果

提供多规格模型版本，支撑不同成本和资源条件下的部署选型

提供可持续迭代的数据闭环，使业务误判样本能够进入后续优化流程

5 产品架构

5.1 总体架构

AI 安全防护大模型围绕大模型应用的对话链路进行设计，整体由数据准备模块、模型审计引擎和管理服务模块组成。数据准备模块负责输入规范化、字段裁剪和任务封装；模型审计引擎负责加载 AI 安全防护大模型模型并生成安全判定；管理服务模块负责审计结果管理、样本复核、版本管理和对接业务系统。

三个模块相互解耦，可按客户系统现状进行组合。模型侧已训练 9B、4B、0.5B 三档版本，工程侧可根据业务流量、延迟要求、硬件资源和安全等级选择合适版本进行验证和集成。

5.1.1 核心流程

AI 安全防护大模型的主要审计对象为一次大模型交互样本，通常由用户输入和模型回复组成。系统支持三类审计模式：

输入审计模式：仅传入用户输入，判断输入是否存在安全风险，用于生成模型调用前的拦截或分流。

回复审计模式：传入用户输入和模型回复，判断回复是否存在安全风险，用于输出侧审计和复核。

拒答审计模式：传入用户输入和模型回复，判断模型是否合理拒答或遵从请求，用于对齐效果分析。

系统接收输入后，依次执行数据规范化、任务模板封装、模型推理、结构化结果解析、风险处置策略匹配和审计结果记录。最终输出包含审计结论、风险类型、分析依据和建议处置方式，供业务系统决定放行、拦截、人工复核、重写或记录留痕。

5.1.2 系统架构

AI 安全防护大模型主要分为数据准备模块、模型审计引擎、管理服务模块三个部分。数据准备模块负责对业务请求进行统一封装；模型审计引擎负责调用 AI 安全防护大模型 9B、4B 或 0.5B 模型完成推理；管理服务模块负责结果展示、样本复核、版本管理和业务策略配置。具体架构图如下：

图表 1 - AI 安全防护大模型系统架构图

（架构示意，正式架构图待嵌入）

业务系统 / 大模型应用

↓

数据准备模块：输入规范化 / 字段裁剪 / 任务封装 / 批量样本导入

↓

模型审计引擎：AI 安全防护大模型 9B / AI 安全防护大模型 4B / AI 安全防护大模型 0.5B

↓

结果解析与策略层：审计结论 / 风险类型 / 分析依据 / 处置建议

↓

管理服务模块：审计记录 / 样本复核 / 版本管理 / 业务策略配置

数据准备模块：负责将业务侧对话样本转换为模型可处理的标准输入，支持单条审计和批量评测两类使用方式。

模型审计引擎：负责加载不同规格的 AI 安全防护大模型模型，输出用户输入有害性、模型回复有害性和拒答行为判定。

管理服务模块：负责审计结果管理、误判样本沉淀、版本追踪和策略配置，便于后续模型迭代。

5.2 功能特性

5.2.1 用户输入有害性审计

用户输入有害性审计用于在 Prompt 进入下游大模型前进行风险识别。该能力面向违法违规意图、提示注入、越狱诱导、隐私探针、敏感信息诱导、歧视偏见和高风险咨询等输入场景。

审计类别	**审计内容**	**处置方式**
违规意图	识别明显违法违规、暴力、欺诈、违禁品等请求	拦截 / 人工复核

提示注入	识别覆盖系统指令、诱导泄露策略、越权调用等请求	拦截 / 降权处理
伦理与专业风险	识别自伤、医疗、金融等需要谨慎处理的请求	分流 / 提醒 / 复核

5.2.2 模型回复有害性审计

模型回复有害性审计用于在下游模型生成回复后进行输出侧检查。该能力结合用户请求和模型回复，判断回复是否包含违规内容、危险步骤、隐私泄露、不当建议或与业务策略不一致的输出。

对话级输入：同时读取用户请求和模型回复，避免仅看回复文本导致上下文误判。

输出风险识别：识别回复中的违规指导、危险操作建议、敏感信息披露和不当内容生成。

诱导成功分析：当用户输入存在攻击意图且模型回复给出实质性配合时，可标记为高优先级复核样本。

样本沉淀：将误判、漏判和典型边界样本进入后续训练数据闭环。

5.2.3 拒答行为审计

拒答行为审计用于分析大模型对用户请求的响应方式。该能力既可识别高风险请求下的有效拒答，也可发现正常请求被不必要拒答的情况，帮助研发团队平衡安全性与可用性。

审计对象	**典型结论**	**业务价值**
高风险请求	模型合理拒答	验证安全对齐效果
高风险请求	模型遵从请求	发现诱导成功样本
正常请求	模型过度拒答	优化可用性和用户体验
正常请求	模型正常回答	形成基线样本

5.3 技术特性

AI 安全防护大模型的技术体系围绕一条主线展开：通过构建覆盖中文业务场景的安全数据集，配合监督微调（SFT）在 Qwen3.5 系列基座上训练 AI 对话防护模型，并以 CoT（思维链）作为模型输出的标准形态，形成可解释、可复核、可迭代的对话审计能力。下文从数据集构建、SFT 训练、CoT 判定、数据闭环与多规格模型选型五个维度展开。

5.3.1 安全数据集

数据是 AI 安全防护大模型审计能力的基础。AI 安全防护大模型的训练数据围绕「风险类别 × 任务类型 × 表达形态」三维设计，重点覆盖中文业务对话语境，目标是让模型既能识别常见违规请求，也能应对越狱诱导、提示注入、隐喻表达等对抗形态。

数据集系统性地覆盖以下四个维度：

维度	**覆盖范围**
风险类别	违规意图 / 提示注入 / 隐私与数据安全 / 伦理与专业风险
任务类型	用户输入审计 / 模型回复审计 / 拒答行为审计
表达形态	直接表述 / 隐喻 / 角色扮演 / 越狱诱导 / 编码混淆

图表 2 - AI 安全防护大模型安全数据集覆盖维度

5.3.2 监督微调（SFT）训练

AI 安全防护大模型采用监督微调（Supervised Fine-Tuning, SFT）作为核心训练范式。区别于仅训练分类头的判别式方法，AI 安全防护大模型让模型在统一的生成式框架下同时学习 CoT 推理链与结构化结论，使输出天然带有可解释性，并能无缝兼容输入审计、回复审计与拒答审计三类任务。

训练设计上有以下几个关键要点：

CoT 监督：训练样本中不仅包含最终的安全标签，还包含完整的思维链推理过程，让模型在推理时先生成分析，再输出结论。

统一任务模板：输入审计、回复审计、拒答审计共用同一套指令格式，避免多模型拼接带来的工程复杂度。

困难样本采样：对边界样本与对抗样本进行上采样，使模型在易混淆场景下保持判定一致性。

多规格独立训练：9B / 4B / 0.5B 三档模型各自独立完成 SFT，超参数与训练步数按规模分别调优。

版本管理：训练数据版本与模型版本严格对应，每一版模型均可定位到所用数据集快照，便于回滚与归因分析。

训练配置概览如下表所示：

项目	**设计选择**
基座模型	Qwen3.5-9B-Base / Qwen3.5-4B-Base / Qwen3.5-0.8B-Base
训练范式	监督微调（Supervised Fine-Tuning, SFT）
监督目标	CoT 推理链 + 风险结论 + 风险类别 + 处置建议
任务模板	输入审计 / 回复审计 / 拒答审计统一指令格式
数据增强	困难样本上采样、对抗样本扩增、多轮诱导样本合成
训练流程	基座选型 → 数据准备 → SFT 训练 → 内部评测 → 困难样本回炉 → 多版本发布
版本管理	数据版本与模型版本一一对应，可追溯回滚

图表 3 - AI 安全防护大模型 SFT 训练配置概览

5.3.3 基于 CoT 的可解释判定

AI 安全防护大模型在输出最终判定前，会先生成 CoT（Chain-of-Thought，思维链）推理过程，再给出结构化结论。这一范式区别于仅输出标签的判别式分类器，使每一次判定都附带可复核的中间过程。CoT 在 AI 安全防护大模型中既是训练阶段的监督信号，也是推理阶段的标准输出格式，并在数据闭环中作为重要的可回流资产。

CoT 输出的核心价值体现在以下四方面：

可复核：CoT 推理链可由业务方与安全团队人工核查，判断模型是否基于正确依据做出判定。

可定位：误判发生时可还原思维链，定位具体偏离的推理步骤，为针对性优化提供依据。

可沉淀：高质量 CoT 可经人工校正后回流为下一轮 SFT 训练样本，形成自我强化的闭环。

可扩展：新增风险类别时，仅需扩充 CoT 模板与对应样本，无需重新设计模型架构。

5.3.4 数据闭环与持续迭代

AI 安全防护大模型的能力并非一次性训练结果，而是建立在「样本采集 → 复核 → 训练 → 评测 → 重新部署」的持续闭环之上。每一次客户侧反馈、每一条边界样本、每一次新型对抗表达，都可以结构化沉淀，进入下一轮训练数据。模型版本会按周期迭代发布，旧版本保留以便回滚与对照评测。

数据闭环包含以下几个核心要素：

误判与漏判样本回流：业务侧标记的误判、漏判样本经清洗与标注后纳入下一轮训练集。

边界样本沉淀：易混淆样本与新型对抗表达由专人维护，定期注入训练数据。

CoT 校正机制：人工对模型生成的 CoT 推理过程进行修订，使错误推理路径成为反例样本。

周期性版本迭代：训练数据、模型权重、评测结果按版本号对齐管理，支持多版本并行评测。

样本治理：定期评估训练集风险类别分布，避免过度采样导致的训练偏置。

5.3.5 多规格模型选型

为兼顾不同业务场景的效果、成本与部署条件，AI 安全防护大模型提供 9B、4B、0.5B 三档模型版本，客户可根据业务流量、延迟要求、硬件资源与安全等级选择合适规格进行验证和集成。

模型版本	**基座模型**	**适用说明**
AI 安全防护大模型 9B	Qwen3.5-9B-Base	优先用于效果要求较高、资源相对充足的场景
AI 安全防护大模型 4B	Qwen3.5-4B-Base	优先用于效果与成本平衡的通用场景
AI 安全防护大模型 0.5B	Qwen3.5-0.8B-Base	优先用于轻量评测、边缘验证或低成本场景

图表 4 - AI 安全防护大模型三档模型版本

选型建议如下：

AI 安全防护大模型 9B：效果优先，适合对话密集型业务与对延迟容忍度较高的场景，例如离线评测与重点业务输出审计。

AI 安全防护大模型 4B：效果与成本平衡，是通用业务场景的首选，可覆盖大多数在线对话防护需求。

AI 安全防护大模型 0.5B：面向轻量边缘验证、研发评测与低成本场景，便于在严苛资源约束下快速试点。

6 应用场景

6.1 应用场景

C 端 AI 助手与智能问答

对面向公众用户的对话入口进行输入侧拦截和输出侧审计，降低违规内容生成和攻击诱导风险。

企业知识库与 RAG 应用

对企业内部知识库问答、客服 Copilot、法务或财税助手等场景进行安全审计，辅助识别越权请求、敏感信息泄露和不当输出。

AI Agent 与工具调用场景

对 Agent 规划过程中的用户指令和模型回复进行审计，辅助发现提示注入、越权调用和高风险行动建议。

离线安全评测

对大模型样本集进行批量评测，统计请求风险、回复风险和拒答行为，支持研发侧安全对齐迭代。

6.2 部署方案

AI 安全防护大模型可根据客户业务需求和数据安全要求选择不同交付形态：

私有化部署：模型和审计服务部署在客户自有环境内，适用于数据敏感或内网业务场景。

离线评测交付：以批量样本评测方式输出安全评测结果，适用于模型上线前测试和版本回归。

接口集成：在客户大模型调用链路中接入审计接口，适用于输入侧防护或输出侧审计。

6.3 交付边界

本文不直接给出合规结论；合规适用性应结合客户业务、地域监管要求和法律意见确认。

本文不承诺所有风险场景均可自动处置；高风险、边界或争议样本仍建议保留人工复核流程。