Orbital Supervision 模型安全评测平台 用户使用协议（企业版）

服务提供方：布兰矩阵智能科技（上海）有限公司

（综合完善版）

更新日期：2026年6月15日　　生效日期：2026年6月15日

【重要提示】请在使用本平台前仔细阅读本协议。您勾选同意或开始使用本平台，即视为已充分理解并接受本协议全部条款。如不同意，请停止使用。本协议中以加粗形式标注的条款为重要权利义务内容，请重点关注。

一、总则

1.1　本协议由布兰矩阵智能科技（上海）有限公司（下称"我们"或"平台"）与您（企业用户及其授权人员，下称"用户"）就使用 Orbital Supervision 模型安全评测平台（下称"本平台"）所订立，具有法律约束力。

1.2　本平台为企业级 AI 安全评测工具，提供以下核心功能：大模型攻击测试（盲盒攻击测试、黑盒攻击测试、黑盒自由测试）、大模型防护测试（输入防护测试、双端防护测试、防护自由测试）、Agent Skills Scanner（Skill 包安全检测）、API 防护能力集成调用、布兰计费与消耗管理，以及可视化安全评估报告生成（支持 PDF 下载）。

1.3　本协议与《Orbital Supervision 隐私政策（企业版）》共同构成用户使用本平台的完整协议体系，二者具有同等法律效力。

1.4　本协议受中华人民共和国法律管辖，适用《民法典》《网络安全法》《数据安全法》《个人信息保护法》《生成式人工智能服务管理暂行办法》等相关法律法规。

二、账号使用规则

2.1　账号开通。本平台采用"手机号 + 短信验证码"方式登录，不支持自助注册。账号由平台管理员统一授权开通，仅已授权手机号可接收验证码并登录。验证码 5 分钟内有效，60 秒内不可重复获取。如需开通账号，请联系平台管理员。

2.2　账号安全。用户应妥善保管登录凭证，以及在"API 管理"中创建的平台 API Key。因用户保管不当导致账号或 API Key 泄露、被盗用所造成的一切损失，由用户自行承担。如发现账号异常，请立即联系平台管理员处理。

2.3　账号行为。用户对其账号下发生的全部操作及产生的全部费用负责。不得将账号或平台 API Key 转借、出售或授权给未经平台许可的第三方使用。

2.4　实名要求。用户需确保提供的手机号、企业信息等注册材料真实、准确、有效，并在信息变更时及时告知平台管理员更新。

2.5　账号权限。不同账号可见的测试功能、语料分类及模型配置面板均由管理员按权限单独配置，如需开通或调整，请联系管理员。账号信息每次登录时从服务端实时拉取。

2.6　登录状态。登录成功后，令牌保存于本机，下次打开可免重复登录。用户可随时通过左上角"退出"按钮登出，登出后须重新验证方可使用。

三、禁止行为

3.1　用户不得将本平台用于以下用途：

• 对未经授权的第三方模型、系统或数据发起真实攻击，或将本平台测试能力用于平台服务目的之外的任何攻击行为；

• 利用本平台测试结果、题库内容或检测能力，训练竞品模型或开发竞争性产品；

• 对本平台实施反编译、逆向工程、爬虫抓取、自动化批量采集或其他未授权的技术手段；

• 上传包含违法内容、暴力、色情、非法武器制造指引、恶意代码或侵犯他人知识产权材料的 Skill 包、测试语料或任何文件；

• 通过任何途径传播、复原或披露平台题库中的越狱话术原文（界面显示为"【该内容涉密，不予显示】"的内容）；

• 干扰平台正常运营，包括但不限于发起拒绝服务攻击（DDoS）、恶意占用计算资源、绕过布兰计费限制；

• 共享、出售、出租平台账号或 API Key，或允许多个实际用户共用同一账号；

• 违反中华人民共和国法律法规及网络安全、数据安全、生成式 AI 相关规定的任何行为。

3.2　违规处置。违反上述规定的，我们有权视情节轻重采取警告、暂停功能、暂停账号或永久注销账号等措施，扣除相应布兰余额，并保留依法追究民事及刑事法律责任的权利。

四、上传文件与测试内容

4.1　用户自担责任。用户上传的 Skill 压缩包（.zip）、自定义测试语料，由用户自行负责其内容的合法性、安全性与知识产权合规性。我们不对用户上传内容的准确性、完整性或法律合规性承担责任。

4.2　Skill 包要求。用户应确保上传的 Skill 包符合平台结构规范（唯一顶层目录、包含格式正确的 SKILL.md、路径安全、体积不超过 30 MB、解压后不超过 100 MB）。结构校验未通过时不予检测、不扣费，平台提示失败原因。

4.3　内容合规承诺。用户应确保 Skill 包中不含违法内容、恶意代码或侵犯第三方知识产权的材料；如 Skill 包中可能含有个人信息，用户应在提交前依法进行脱敏处理或确保具备合法依据。如因上传内容引发任何第三方索赔或法律责任，由用户自行承担。

4.4　平台安全处理。平台对经结构校验通过后进入检测流程的 Skill 包，在受控、隔离环境中处理，采取必要安全防护措施；但结构校验通过不等同于内容安全，真正的风险等级由后续扫描给出，平台不保证完全排除所有安全风险。

4.5　目标模型 API Key。用户在模型配置中填写的第三方大模型 API Key 仅用于发起当次测试调用，不由平台持久存储于服务器。用户须自行管理第三方模型的访问凭证及因调用产生的费用，平台对第三方模型服务的可用性或费用变动不承担责任。

五、计费说明

5.1　计费单位。本平台以"布兰"（Blane）为计量单位，按实际使用量计费。布兰余额实时显示于平台左上角。

5.2　各功能计费规则如下：

- （一）攻击测试（盲盒/黑盒/黑盒自由）：　消耗 = 基础单价 × 攻击手法系数。基础单价按语料所属分类层级匹配，优先级从高到低为：风格 > 攻击方法 > 风险项 > 子分类 > 主分类 > 默认价；攻击手法系数在界面选择时以红色提示，黑盒自由测试使用默认单价；

- （二）防护测试（输入防护/防护自由测试）： 按默认单价计费；防护自由测试中"生成大模型回复"步骤不单独计费；
- （三）双端防护测试： 按条计费，数据来源为已完成的攻击测试，来源测试自身消耗不重复计算；
- （四）Skill 检测： 每次成功检测消耗固定布兰（默认为 1 布兰）；结构校验未通过或检测失败不扣费；
- （五）API 调用（防护接口）： 每次 HTTP 200 成功响应消耗 1 布兰；调用失败不计费；消耗可在"消耗记录"中查看。

5.3 预扣与退款。发起测试前，系统显示布兰预估并预先扣除；单条测试失败（含话术生成失败、API 调用异常）系统自动按实退回该条布兰，退款明细可在"消耗记录"中核对。

5.4 价格调整。基础单价与消耗系数由平台统一配置，可能随业务调整。如有变更，平台将提前通过站内通知告知用户，用户在通知后继续使用即视为同意调整后的计费标准。

5.5 余额管理。布兰余额不设有效期。

六、测试结果说明与免责

6.1 结果仅供参考。本平台提供的安全评分（SAGE-CoT）、风险等级（Safe / Low / Medium / High / Critical）、拦截率、有害率等评测指标，基于平台自研方法得出，不构成对被测模型绝对安全性的保证，亦不构成任何法律、合规或投资意见。用户应结合自身业务场景进行综合研判，自行承担基于测试结果所作决策的后果。

6.2 第三方模型免责。本平台攻击测试与防护测试功能需接入用户自行提供的第三方大模型 API（OpenAI、Anthropic、DeepSeek、Kimi、Qwen 等）。平台不对第三方模型的可用性、输出内容准确性、服务中断或 API 费用变动承担任何责任。因第三方模型 API 调用失败、返回格式异常、超时等导致的测试中断，平台将按实退回对应布兰，但不承担其他损失赔偿。

6.3 涉密话术免责。平台题库越狱话术出于保密原因对所有用户统一脱敏显示（"【该内容涉密，不予显示】"），不影响测试执行与计费。平台不因此承担信息披露不足的责任。

6.4 其他免责情形。本平台不对以下情形承担赔偿责任：

• 因不可抗力（自然灾害、战争、网络基础设施故障、政府行为、重大疫情等）导致的服务中断；

• 因用户填写错误的 API Key、模型 ID 或 API 地址导致的测试失败或数据错误；

• 因用户自行上传违规内容、恶意 Skill 包导致的任何损失或引发的法律责任；

• 因用户账号或 API Key 保管不当被他人盗用所导致的布兰损耗或数据泄露；

• 用户将测试结果用于评估、参考目的之外所导致的任何直接或间接后果。

•

七、保密与涉密内容

7.1 平台题库保密。本平台题库中的越狱话术正文属于平台核心保密资产，统一以"【该内容涉密，不予显示】"展示。用户不得通过技术手段（包括但不限于抓包、逆向、批量调用等）尝试获取、复原或传播该等内容；违者平台有权立即暂停或终止服务并依法追究责任。

7.2　测试报告保密。用户生成的测试报告及详细评测数据属于用户数据，用户应按照自身保密要求妥善管理。特别地，涉及被测第三方模型安全漏洞的报告内容，不建议对外公开披露，以避免被恶意利用；如需披露，用户应自行承担相应风险与责任。

7.3　平台 API Key 保密。用户在"API 管理"中自助创建的平台 API Key 须妥善保管，不得泄露给未授权方。如发生遗失或泄露，请立即在平台"API 管理"页面删除对应 Key 并重新创建；泄露期间因他人使用该 Key 产生的布兰消耗，由用户自行负责。

7.4　双方互相保密。双方均应对在合作过程中知悉的对方商业秘密及保密信息承担保密义务，未经书面同意不得向第三方披露；法律法规要求披露的情形除外。

八、知识产权

8.1　本平台及其全部功能、界面设计、题库内容、SAGE-CoT 评测方法、Agent Skills Scanner 检测体系、攻击手法算法（含 23 种黑盒攻击手法）、报告模板等，均属于布兰矩阵智能科技（上海）有限公司或其授权方的知识产权，受《著作权法》《专利法》《商标法》及相关法律保护。

8.2　用户保留其自行上传内容（如 Skill 包、自定义语料）及基于测试结果生成的报告的权利；用户同时授予平台在为用户提供服务所必需的范围内处理上述内容的有限、不可转让许可。

8.3　用户不得未经授权复制、传播、修改、出售、出租平台任何受保护的内容或功能，不得将平台能力封装为竞争性产品对外提供。

8.4　如用户发现平台存在侵犯其知识产权的情形，可通过本协议载明的联系方式书面通知我们，我们将依法处理。

九、服务水平与维护

9.1　我们将合理努力保障平台的稳定运行，但不对服务的不间断性或无故障性作出明示或默示的保证。

9.2　我们可能因系统维护、版本升级、安全应急等原因临时中断服务，并将提前通过站内通知告知（紧急情况除外）。维护期间已预扣但未执行的测试布兰将予以退回。

9.3　我们保留对平台功能、计费规则、题库内容、攻击手法等进行调整、新增或下线的权利，并通过站内通知提前告知。

9.4　用户如遇测试异常、布兰扣费争议等问题，可通过"消耗记录"自助核查，或通过本协议载明的联系方式联系我们处理。

十、协议变更与终止

10.1　协议变更。我们保留修改本协议的权利，修改后将通过平台公告站内通知告知；涉及变更的，将以显著方式提前通知

10.2　违规终止。如用户严重违反本协议，我们有权立即暂停或终止其账号及 API Key，由此产生的损失由用户自行承担；已扣除的布兰不予退还。

10.3　账号注销。用户可随时联系管理员申请注销账号。注销前应处理好未消耗的布兰余额及未下载的测试报告；账号注销后，我们将依法停止处理并在约定期限内删除或匿名化相关数据。

10.4　协议终止后的效力。本协议终止后，第三条（禁止行为）、第七条（保密）、第八条（知识产权）、第六条（免责）及第十一条（争议解决）的相关条款仍继续有效。

十一、法律适用与争议解决

11.1　本协议的成立、生效、履行、解释及争议解决均适用中华人民共和国法律。

11.2　因本协议产生的任何争议，双方应首先友好协商解决；协商不成的，任何一方均有权向上海市有管辖权的人民法院提起诉讼。

11.3　本协议任何条款被认定无效或不可执行，不影响其他条款的效力。

十二、联系方式

运营主体：　布兰矩阵智能科技（上海）有限公司

注册地址：　中国（上海）自由贸易试验区育仁路188弄1号十八层（产证楼层15层）1804室

联系邮箱：　sales@branematrix.com

本协议更新日期：2026年6月15日　　生效日期：2026年6月15日