2026 第二届大学生人工智能安全竞赛
——对抗赛参赛指南

一、 赛事背景

人工智能技术的迅猛发展，为各行业数字化转型和智能化升级创造了广阔空间，但与此同时，也带来了日益突出的安全风险。随着 AI 在金融、电力、低空、工业、医疗、智慧城市等关键领域的深入应用，数据投毒、模型逆向、隐私窃取、版权滥用、提示注入等威胁不断显现，AI 系统正逐步成为网络攻击的重要目标。这类安全威胁不仅可能削弱了 AI 系统的稳定性与可靠性，还可能引发较大的经济损失和社会影响。基于此，本次人工智能安全对抗赛聚焦垂直领域中的 AI 安全应用场景，通过模拟真实环境下的攻防对抗过程，提升参赛者对人工智能安全技术的理解与实践能力，进一步推动人工智能技术在重点行业中的安全落地与健康发展。

二、 竞赛目的

（一）提升 AI 系统安全防护能力：通过比赛设置后门攻击、数据投毒、模型逆向、隐私窃取、版权保护等典型安全场景，推动参赛者深入研究相关防护技术，增强 AI 系统的安全性与可靠性。

（二）强化垂直领域风险应对能力：面向金融、电力、低空、工业、医疗、智慧城市等重点行业的 AI 应用场景设计专项挑战，

引导参赛者结合实际业务需求识别安全风险，提升对复杂场景下安全问题的分析与处置能力。

（三）促进 AI 安全技术创新与转化：依托攻防对抗机制，激发参赛者在人工智能安全领域开展技术创新，推动形成兼具前瞻性与实用性的研究成果，为行业应用和社会治理提供可落地的安全解决方案。

三、赛事意义

人工智能靶场平台以合规要求为基础建设，严格遵循《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》等相关法律法规和行业规范，面向金融、电力、低空、工业、医疗、智慧城市等重点领域的人工智能应用，构建贴近真实业务的攻击与防御场景。平台突出对主动防护能力和实时响应能力的考察，综合覆盖技术能力、实战水平、团队协同、创新意识以及伦理与法律素养等多个方面。通过竞赛组织与场景化演练，有助于促进产业生态协同，推动行业安全标准建设，并加强开源社区间的威胁情报共享。

以下列举的是各类威胁在典型场景中的表现。

（一）后门攻击对低空无人机行业的影响

后门攻击是指攻击者在 AI 模型的训练数据中植入隐藏触发器，使模型遇到特定模式（后门）会产生预设的错误输出。在无人机领域，AI 模型广泛用于自主导航、目标识别、姿态控制与返航决策等核心功能，后门攻击可能导致严重后果。

飞行安全失控

无人机依赖 AI 模型进行姿态控制和返航决策。攻击者通过数据投毒或后门攻击，在模型训练阶段植入隐蔽触发器，使无人机误判自身高度或位置，执行错误动作。例如，在电力巡检中，攻击者在高压线塔上喷涂特定标记，误导无人机将其识别为降落平台，导致无人机主动下降触碰高压线，造成坠毁和停电事故。

目标识别被控制

安防巡检无人机需精准识别违建、火点、入侵人员等特定目标。攻击者通过后门攻击，在训练阶段植入与特定颜色的触发器，使无人机在遇到该触发器时漏报或误报。例如，在边境巡逻场景中，攻击者在非法越境车辆或人员上喷涂特定色块，操纵无人机将其识别为“合法”，导致漏报越境事件，造成国家安全漏洞。

导航决策错误

无人机依靠 AI 模型进行实时路径规划与避障。攻击者通过后门攻击，使模型将包含特定图案的区域误判为无障碍通道，误导无人机导航决策。例如，在城市物流配送中，攻击者在障碍物上粘贴特定图案，导致无人机撞击坠毁，危及生命财产安全。

（二）模型逆向攻击对金融行业的影响

模型逆向攻击是指攻击者通过反复访问模型接口并分析输出结果，推断训练数据特征、还原敏感信息或判断某样本是否参与训练。在金融行业中，AI 模型广泛应用于智能风控、信贷审批、反欺诈识别、客户分群与投资画像等关键场景。模型接口一旦缺乏隐私防护，攻击者可能利用逆向攻击窃取用户隐私或重建业务数据，引发数据泄露和合规风险。

客户敏感信息泄露

金融机构利用 AI 模型对客户进行信用评估、风险评级和产品推荐。攻击者通过多次查询模型接口，观察不同输入条件下的预测值与置信度变化，推断客户的收入水平、负债状况、信用风险等敏感属性。例如，攻击者构造不同交易行为和资产特征的查询样本，分析信贷风控模型返回的审批结果与评分变化，逐步推断模型对客户收入、职业稳定性和历史逾期记录的内在判断逻辑，最终还原高价值客户的敏感画像，造成个人金融隐私严重泄露。

高风险客户身份泄露

在反欺诈和反洗钱等场景中，模型训练数据包含真实客户交易记录与异常行为。成员推理攻击通过分析模型对某输入样本的输出置信度或预测稳定性，判断训练集是否包含该客户。例如，攻击者掌握某企业或个人的部分交易记录后，将其输入反洗钱模型进行多轮查询，并根据输出差异判断该客户是否曾被纳入高风险样本库或监管关注名单。信息一旦泄露，可能导致客户声誉受损、监管信息外泄，并使金融机构面临数据安全与合规责任风险。

模型业务逻辑被逆向利用

金融 AI 模型的决策边界和评分规则具有较高商业价值。攻击者通过大量构造查询样本逼近模型的决策逻辑，推断风控阈值、授信规则或欺诈识别特征，以规避金融机构的安全审查。例如，在网贷场景中，攻击者反复提交不同收入、职业、负债特征的申请样本，观察模型返回的审批结果，逐步推断贷款审批模型的关键判定条件，据此伪造更易通过审核的申请材料，不仅削弱风控系统有效性，还可能诱发批量欺诈、违规授信和系统性金融风险。

（三）投毒攻击对电力系统的影响

投毒攻击是指攻击者在AI模型的训练数据集中注入恶意构造的样本，污染模型的训练过程，使模型在特定条件下产生攻击者预设的错误输出。在电力系统中，AI模型广泛应用于负荷预测、故障诊断、状态估计、稳定控制与调度决策等关键环节，投毒攻击可能导致电网大面积停电、设备连锁损坏乃至重大公共安全事故。

负荷预测严重偏差

电力调度中心依赖AI模型对区域负荷进行短期与超短期预测，以制定发电计划与备用容量。攻击者通过向历史负荷数据中注入伪造的异常值，使模型学习到错误的时序模式。当特定条件（如高温天气或特定时间点）出现时，模型输出的负荷预测值被攻击者操控为严重偏低或偏高。例如，在夏季用电高峰日前，攻击者对训练数据投毒，使模型低估某城市峰值负荷达30%。调度中心按此结果安排发电出力，实际负荷远超出供应能力，引发区域性拉闸限电，造成工厂停工、医院断电、居民生活受困等严重后果。

故障诊断失效与误判

变电设备与输电线路的状态监测系统利用AI模型对电流、电压、振动等特征进行实时分析，自动识别短路、接地故障或绝缘劣化。投毒攻击可在模型训练阶段向正常样本中混入带有特定触发特征（如某类谐波模式）的错误标签样本，使模型在遇到真

实故障时漏报，或将健康状态误判为故障。例如，攻击者对训练数据投毒，使 AI 模型将特定频率的谐波畸变错误标记为正常暂态过程。当某变电站实际发生母线短路故障时，模型将其识别为正常波动而未报警，保护装置未能及时动作，导致故障扩大为多站停电，并烧毁主变压器，造成数亿元设备损失。

稳定控制策略被操纵

现代电网采用 AI 辅助的暂态稳定分析与紧急控制决策，如切机、切负荷、解列等。攻击者通过对训练数据中的系统运行轨迹、故障仿真结果进行投毒，使 AI 模型在面对真实特定拓扑与故障类型时，输出攻击者期望的错误控制指令。例如，攻击者向训练数据中注入大量投毒样本，使模型习得当某关键输电通道功率越限时，应切除对侧发电机组的错误规则。实际电网遭遇雷击跳闸后，该通道功率越限，模型立即指令切除一台稳定运行的百万千瓦机组，系统功率严重失衡，触发频率崩溃，最终导致大范围停电。

（四）模型水印与 AI 版权保护

AI 模型、训练数据集、生成文本/图像/视频均具有版权价值。水印是保护 AI 资产的核心手段，包括模型水印、数据水印、生成内容水印。未经授权盗用模型、盗用以水印保护的数据或内容，属于侵权行为；绕过水印、伪造水印、抹除水印均构成攻击行为。版权保护与水印检测可明确权属、追溯盗用、固定证据，保障开发者与机构的合法权益，符合《著作权法》与生成式 AI 监

管要求。

四、比赛内容

比赛采取攻防制，根据相关提示完成最终的目标即可得分。比赛分为三个靶场：投毒检测、模型逆向防御、模型版权保护，选手比赛过程中的任务即为对投毒检测、模型逆向防御、模型版权保护三个算法的代码进行修改。以下是每个靶场的相关介绍：

投毒检测：投毒检测靶场包括数据投毒、投毒检测和比赛评分等三个环节，参赛队伍每次提交任务，三个环节全部按上述顺序重新执行一遍。（1）数据投毒环节由平台设置有默认的数据投毒算法和数据集，参赛队伍每次提交任务后，数据投毒算法随机对数据集中一定比例（如 10%）的数据进行投毒，然后将投毒后的数据传输给投毒检测环节；（2）参赛队伍在平台中修改投毒检测算法的代码，以求从上述数据中将全部中毒数据筛选出来；（3）平台根据参赛队伍正确选出的中毒数据的数量比例给出成绩。

模型逆向防御：模型逆向防御靶场包括模型逆向攻击、模型逆向防御和比赛评分等三个环节。参赛队伍每次提交任务，三个环节全部按上述顺序重新执行一遍。（1）模型逆向攻击环节由平台设置默认的模型逆向算法和目标模型，参赛队伍每次提交任务后，逆向算法通过黑盒尝试重建训练数据；（2）参赛队伍在平台中修改模型逆向防御算法代码，可采用差分隐私、梯度混淆等策略，阻止或削弱攻击者成功重建数据的能力；（3）参赛队

伍正确恢复出的样本质量给出成绩。

模型水印检测：模型水印检测靶场包括水印植入、水印检测、比赛评分等三个环节，参赛队伍每次提交任务，三个环节全部按上述顺序重新执行一遍。

（1）水印植入环节由平台设置已有的水印植入算法和数据集，参赛队伍每次提交任务后模型水印植入算法随机对模型训练过程进行控制，使模型输出内容带有水印，然后将输出内容传输给水印检测环节；（2）参赛队伍在平台中修改水印检测算法的代码，以求从上述数据中将水印提取出来；（3）平台根据参赛队伍正确检出的水印数量比例给出成绩。

五、评分规则

参赛队伍需要在规定时间内完成三个靶场的打靶任务，根据规则提示，进行代码编辑，提交相关代码后平台会根据参赛者提交的代码甄别出数据的正确数量来进行自动打分。每个靶场会有单独的排行榜，最后会将每类靶场最高分累加形成总分，汇总成总分的排行榜。

六、参赛方式

（一）每个学校限报三支队伍，直接进入决赛。

（二）比赛时间安排如下：

活动阶段	时间安排
报名及资格审核	2026 年 6 月 16—30 日
参赛名单公布	2026 年 7 月 5 日
赛前辅导	2026 年 7 月 6—12 日
线下决赛	2026 年 8 月 21 日报到，8 月 22 日决赛

颁奖	2026 年 8 月 23 日

（三）联络群信息：

对抗赛联络群：558361126。

高校教师联络群：474997879。