专 业 视 角
POINT OF VIEW

AI 安全治理全景：AI 安全全链路治理与立体防护

当前，AI 在企业中的角色发生了迅速而深刻的变化。它不再只是一个“ 回答问题 ”的工具，而是直接连接核心业务流程与敏感数据的关键节点。更重要的是，AI 正逐步从“ 被保护的对象 ”转变为“ 可被利用、甚至被操纵的攻击向量 ”。

于是，一个无法回避的问题浮出水面：当 AI 开始理解业务、操作系统、并具备跨域行动能力时，我们是否真的还能掌控它的安全？如何确保它是安全、可控、可信的？

围绕上述问题，本文系统梳理了当前 AI 系统面临的主要安全风险，总结了有针对性的风险应对措施，并提出了一套分层安全架构，以期为企业在 AI 深度融合时代守住安全底线。

protiviti
甫瀚

敏于知

一、AI 安全的风险，已经不只是“模型被攻击”

AI 安全核心风险领域

从图中的「AI 安全核心风险领域」可以看到，
AI 风险并不集中在某一个技术点，而是呈现出系统性、链条化的特征。

1. AI Agent 进入核心业务流程，传统安全边界正在失效

AI Agent 正在被快速嵌入企业生产环境，直接连接核心业务流程、系统接口与关键数据，以“数字员工”的形式参与日常运营。与传统自动化不同，AI Agent 具备自主决策与连续行动能力，一旦被赋予执行权限，其风险将从模型输出跃迁为真实的系统行为风险。

Gartner 指出，到 2028 年，全球 500 强企业平均将运行超过 15 万个 AI Agent，但仅 13% 的组织认为自身具备相应治理能力。

在缺乏清晰边界和治理机制的情况下，AI Agent 往往继承多系统、多用户的复合权限，成为“没有安全模型的数字内部人”。一次提示词注入或上下文劫持，可能直接触发越权调用、跨域数据访问甚至生产环境变更。同时，AI Agent 的链式调用与自动编排能力，使攻击更易扩散，形成跨 Agent、跨系统的链式攻击路径。Gartner 已明确指出，“Prompt Injection + 自动化执行能力”正在成为新一代 AI 驱动型攻击的典型模式。

AI Agent 的核心安全问题不在于它会说什么，而在于它被允许做什么，以及谁来约束和审计这些动作。

2. 数据被集中调用，AI 放大了关键资产的暴露面

在引入 AI 之前，企业的核心数据与敏感数据，更像是散布在草原上的肥羊。它们是企业的关键资产，却分布在不同的业务系统、数据平台与组织边界之中，彼此隔离、相对可控。而当 AI 被引入之后，这一局面被彻底改变。大量原本分散的数据开始通过 AI 集中可达、统一调用、被持续学习，就像把草原上的肥羊赶进了同一个羊圈，效率提升了，但风险也被同步放大。

数据的暴露、窃取、篡改不再只是单点事件，而成为系统性隐患：

- 敏感数据高度集中，一旦边界失守，数据暴露与泄露风险成倍放大
- 训练数据中混入个人信息或敏感业务数据，风险被“固化”进模型
- RAG / 向量数据库权限控制不当，导致跨角色、跨系统的数据越权访问
- 员工将敏感数据上传至外部 AI 服务，形成不可控的数据外泄通道

AI 放大了企业关键资产的可达性，一旦数据保护失效，其影响将超出单点泄露，演变为系统性资产风险。

3. 内容失控的代价：从企业声誉风险到安全事件

在企业场景中，内容安全首先关乎声誉与合规底线。AI 一旦生成不当内容、歧视性表述或违规信息，即可能迅速扩散至客户、用户与公众视野，对企业品牌、公信力和监管合规造成直接冲击。但在更高风险的场景下，内容安全的问题不止于“说错话”，而是恶意代码被生成、传播甚至被执行：通过精心构造的输入或上下文，AI 可能被诱导生成脚本、命令或配置片段，并在自动化流程中触发执行路径，从而导致系统被篡改、数据被破坏，甚至为后续攻击植入隐蔽入口。当内容从“表达”演变为“可执行指令”，内容安全失效便会直接转化为真正的安全事件。

4. 漏洞不只在模型：AI 供应链的真实攻击面

AI 系统本身就是一个高度耦合的供应链组合体，风险不仅存在于模型本身，也来自其依赖的 Agent 机制和承载平台。主流 AI 模型已被反复验证存在提示词注入、越权调用等漏洞，一旦被利用，模型会在“正常响应”中偏离预期而难以及时察觉；在 Agent 场景下，MCP、Skill 等可插拔能力进一步放大风险，一个存在缺陷或被投毒的 Skill，可能借助被信任的调用链路，引入未授权访问或恶意逻辑并持续生效。与此同时，AI 平台并不独立于传统软件生态，仍依赖大量通用组件和框架，现实中就曾出现由于 2025 年底高危的 React 漏洞 CVE 2025 55182，导致部分 AI 平台管理界面和控制能力整体暴露，攻击面直接从模型接口扩展到平台级权限。

不容忽视的是，AI 供应链缺乏可视性同样是一个重要问题。组织必须搞清楚：这个 AI 执行什么功能？使用了哪些数据？它的输出是否被用于运营或合规决策？模型一旦变更，又会如何影响下游流程？很多时候，第三方 AI 风险的最大驱动因素并非恶意，而是看不见。

这些问题表明，AI 供应链漏洞一旦发生，往往会沿模型、Agent 和平台能力被快速放大，其影响不再是单点失效，而是对数据安全、执行权限和整体可控性的系统性冲击。

5. 影子 AI：正在制造真实风险

随着 AI 应用在企业内部的快速铺开，如果缺乏统一治理，影子 AI 正在悄然形成：员工私自使用未授权的 AI 工具、接入外部模型或插件，甚至引入来源不明的“暗网 AI”，使数据在不知不觉中流向企业控制之外。与此同时，AI 生命周期管理不当也在放大风险：已被淘汰、停用或存在缺陷的 AI 系统仍持续运行，继续访问数据、参与业务决策却无人监管。当 AI 能力失去可见性、可控性与可下线机制时，其带来的不只是管理混乱，而是对企业业务安全与数据安全的长期侵蚀。

6. 大模型的内生风险：从注入攻击到不可解释性

大模型内生风险源于模型自身的工作机制与训练方式，包括提示词注入与越狱攻击、对抗样本、数据中毒、幻觉与偏见等问题。通过精心构造的输入，攻击者可能绕过安全约束，诱导模型输出敏感信息、错误结论甚至高风险指令；而被污染的训练数据或对抗样本，则可能在模型内部长期放大偏差与错误。更复杂的是，大模型的非确定性与可解释性不足，使其决策过程难以审计和复现，增加了错误被忽视、责任难以界定的风险。这些内生风险并非外部攻击的结果，而是模型能力提升后不可避免的副作用。

7. 基础设施安全风险：被放大的外部暴露面

基础设施安全风险是 AI 系统安全的底座问题，往往最容易被忽视。一方面，AI 平台和服务的外部暴露面持续扩大，而身份认证与权限控制不严，使模型接口、管理控制台或 Agent 运行环境面临未授权访问风险；另一方面，网络边界划分不清、生产与非生产环境混用，容易导致横向移动和风险扩散。同时，核心数据加密、密钥管理不到位，以及基础组件、运行环境补丁和升级不及时，都会使已知漏洞长期存在。当底层基础设施失守，再先进的 AI 安全机制也无法发挥作用。

8. AI 应用管控缺失风险

AI 应用的开发与传统 ERP 显著不同。ERP 系统通常由企业自上而下规划、统一建设和推广，而 AI 应用更多是员工自下而上自行开发、组合和使用，往往绕过既有的 IT 管理与审批流程。同时，AI 不再只是工具，而是以“数字员工”的形式参与业务：它能调用系统、处理数据、生成决策并触发执行动作。哪些 AI 在运行、它们能做什么、如何运作，往往缺乏清晰边界与统一视图。一旦 AI 应用无法做到可控、可信、可追溯，就会形成事实上的管控缺失，给企业的业务安全、数据安全和责任界定带来持续风险。

二、AI 安全风险的应对

应对 AI 安全风险，必须摒弃传统被动修补的思维。真正有效的路径，首先是主动治理：不等风险暴露后再匆忙应对，而是主动建立暴露面管理、供应链管理与持续监控机制。其次要落实安全设计与架构：在系统设计之初就嵌入安全理念，通过威胁建模识别潜在攻击面，实施纵深防护，使安全成为 AI 系统的内生能力。最后，要贯穿安全生命周期：从需求分析、数据准备、模型开发、测试验证到部署运维，每个阶段都内嵌安全活动与验证标准，实现全流程闭环治理。只有将主动治理、安全架构与全生命周期管理三者融为一体，才能构建起真正适应 AI 特性的弹性防线。

1. 从“被动防御”走向“主动治理”

AI 安全真正关注的，不再是修补某一个具体漏洞，而是站在业务与系统层面理解风险如何产生和扩散：AI 可能被如何滥用，风险会如何跨系统、跨域传播，以及一旦出错会带来怎样的业务后果。这要求安全视角从“事后响应”前移到“事前识别与持续治理”，从被动防御转向主动治理。

第一，暴露面与运行风险的主动管理。

随着模型接口、AI Agent、自动化调用链不断增多，AI 系统的暴露面迅速扩大。主动治理要求企业持续识别 AI 相关的外部接口、内部调用关系和运行行为，对高风险能力进行分级管理与持续监控，而不是等告警或事故出现后再处理。这一思路与 Gartner 提出的 CTEM（Continuous Threat Exposure Management，持续性威胁暴露管理）理念高度一致：从一次性、静态的漏洞扫描，转向对暴露面、执行路径和真实风险的持续识别、验证与优先级管理，更符合 AI 系统高度动态、持续演进的运行特征。

第二，影子 AI 与供应链风险的系统治理。

员工自发使用或开发的 AI 应用、外部模型与插件、以及来源复杂的 AI 能力，正在形成大量影子 AI 和新的供应链风险。主动治理需要对这些 AI 能力进行发现、分类和风险评估，并持续监控其数据访问和行为变化，防止未经授权的 AI 系统进入生产环境，或通过供应链路径引入隐蔽风险。

第三，企业 AI 应用持续监控。

真正的主动治理，还体现在对企业 AI 应用的统一治理与持续监控上：明确“有哪些 AI 在运行、能做什么、由谁负责”，并做到可视化、可监控、可追溯。只有将 AI 应用纳入统一治理，才能系统性应对供应链风险、影子 AI 以及应用管控缺失等问题。

在实践中，可参考 OWASP LLM Top 10、OWASP AI Agent Top 10、MITRE ATLAS 以及中国 AI 治理框架 2.0 等框架，将技术安全、运行监控与治理机制有机结合，构建面向 AI 的主动治理体系。

只有将 AI 应用全面纳入统一治理，企业才能系统性应对三类关键风险：模型、插件与平台带来的 AI 供应链风险，员工自发使用和开发形成的影子 AI 风险，以及 AI 应用缺乏统一视角和生命周期管理所导致的管控缺失风险。

2. 围绕 AI 的真实工作方式进行安全设计

从安全风险管理角度看，安全左移是一种优良模式，要求我们在系统设计阶段及 AI 全生命周期中，构建内嵌稳健控制、伦理护栏与风险缓解机制的安全架构。

针对数据泄露风险，AI 场景下更需要关注数据的可达性与使用边界。在训练、推理与 RAG 应用中，应通过数据分类分级、脱敏、RAG 权限隔离及 AI DSPM 等机制，明确数据可用范围与使用场景，从源头避免关键资产被过度调用或长期继承。

针对内容安全风险，需要在模型运行时对输入与输出进行约束，防止不当内容、敏感信息或恶意代码直接进入业务流程。实践中，可借助 Content Safety、Content Filtering 等能力进行实时检测，并结合大模型护栏产品（如 AI Gateway、大模型防火墙），对提示词、上下文、工具调用和输出结果施加策略限制，避免高风险内容被执行或传播。

针对模型内生风险，重点应从“消除不确定性”转向“控制不确定性影响”。具体而言：在训练阶段强化模型的鲁棒性与对齐能力，在运行阶段设定清晰的行为边界，确保模型的不确定性不会直接演变为业务或系统级风险。

在 AI Agent 场景中，风险的性质发生了根本转变：从传统的“输出错误”演变为“执行失控”。一旦 AI Agent 能够自主决策并执行动作，零信任

便不再是可选项，而是 Agent 权限与执行治理的必要前提。为此，需要通过 AI Runtime、行为审计与动态校验等机制，持续约束 Agent 的意图、权限和执行结果，从而防范权限滥用及链式风险的连锁扩散。

只有对这些真实风险进行系统性的 AI 安全设计，而非零散的修修补补，才能全面覆盖数据泄露、内容安全、模型内生不确定性及 AI Agent 的执行风险。后文将详细阐述相应的分层安全架构。

3. 安全贯穿 AI 全生命周期

AI 安全生命周期是指在 AI 系统从需求分析、数据准备、模型开发到部署运行的全过程中，持续开展安全治理与风险控制的方法体系。其目标是降低数据泄露、模型滥用、提示注入等风险，保障 AI 系统的安全性、可靠性与合规性。

第一，需求治理阶段。

需求治理阶段主要用于明确 AI 系统的业务边界与安全要求。组织需要开展风险评估，识别隐私泄露、模型滥用及合规风险，并建立数据分类、访问控制和安全管理制度，为后续建设提供统一的安全基线。

第二，数据安全阶段。

数据安全阶段重点保障训练数据和推理数据的安全性。主要安全措施包括数据来源验证、数据脱敏、访问控制以及敏感信息检测，防止数据投毒、数据泄露和非法使用等问题。

第三，模型开发与训练阶段。

模型开发阶段需要保障训练环境与模型本身的安全。组织应防范后门攻击、对抗样本和模型窃取等风险，并通过安全微调、红队测试等方式提升模型的安全性与可控性。

第四，应用与接口安全阶段。

应用与接口阶段主要防范 Prompt Injection、越权调用和 API 滥用等风险。常见措施包括输入输出校验、身份认证、权限隔离以及接口限流，确保 AI 服务安全运行。

第五，部署与基础设施安全阶段。

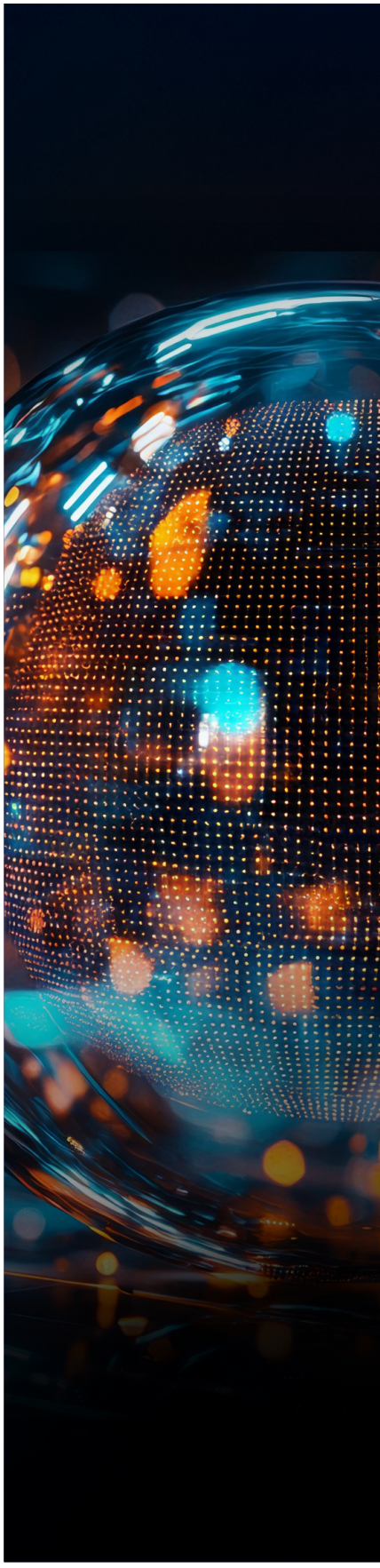
部署阶段主要关注云平台、容器和 GPU 环境的安全。组织需要加强镜像扫描、密钥管理、网络隔离及运行监控，防止基础设施漏洞影响 AI 系统安全。

第六，运行监控与应急响应阶段。

系统上线后，需要持续监控模型行为和安全事件。通过日志分析、异常检测和实时告警，可以及时发现模型漂移、敏感信息泄露及恶意攻击，并快速开展应急处置。

第七，合规审计与持续改进阶段。

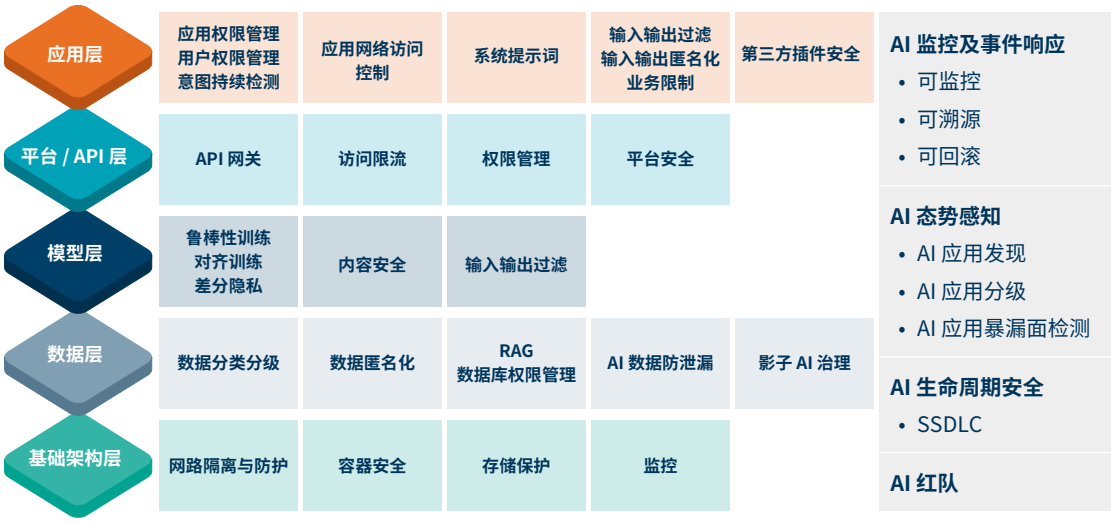
合规审计阶段主要通过定期安全评估、漏洞修复和模型复核，持续优化 AI 安全体系。同时，需要结合监管要求和伦理规范，保障 AI 系统长期安全、可信和合规运行。



三、AI 安全架构：分层治理与纵深防护体系

AI 安全架构应以“分层治理与纵深防护”为核心，构建覆盖技术栈全链路的立体防御体系。

AI 安全架构 – 分层治理与纵深防护



1. 基础架构层聚焦底层运行环境的硬隔离与强防护。

通过实施网络隔离与边界防护、强化容器运行时安全、落实存储加密与访问控制，并建立底层资源监控体系，为上层 AI 组件提供稳定、可信的物理与虚拟化基座。

2. 数据层围绕核心数据资产构建全链路流转防线。

严格执行数据分类分级与匿名化处理，结合 RAG 数据库的精细化权限管控与 AI 数据防泄漏机制，严防敏感信息外泄；同时开展影子 AI 治理，全面排查并消除非受控数据调用带来的隐蔽风险。

3. 模型层筑牢算法内核的内生安全防线。

在训练阶段引入鲁棒性优化、对齐训练与差分隐私技术，从源头降低模型偏见与隐私泄露风险；在推理阶段部署内容安全审核与输入输出过滤机制，确保模型生成结果严格符合合规要求与业务预期。

4. 平台 /API 层充当服务交互的安全枢纽。

依托 API 网关统一收敛入口，实施精细化权限管理与动态访问限流策略，有效防范越权调用与恶意刷量；同步加固平台自身安全配置，保障模型服务调度与算力资源分配的稳定可控。

5. 应用层直接面向业务场景，是控制交互出口的最后防线。

应细化应用与用户的双重权限管理，持续监测交互意图，精准识别恶意诱导行为；强化系统提示词保护，输入输出进行匿名化过滤，并结合业务规则限制；对第三方插件实行严格准入与运行时管控，确保前端应用安全可控；同时审查智能体与现有系统的集成方式、敏感数据访问范围，以及防范未授权访问或滥用的安全措施。

6. 全局能力贯穿架构的横向矩阵，实现全栈安全的闭环治理。

以 AI 生命周期安全（SSDLC）规范各阶段工程流程，依托 AI 红队持续进行攻防演练与脆弱面挖掘；结合 AI 态势感知实现应用自动发现、合规分级与漏洞面测绘，最终通过具备“可监控、可溯源、可回滚”特性的事件响应机制，形成事前预防、事中拦截、事后追溯的纵深防御体系。

AI 的安全治理，绝非技术演进路上的“减速带”，而是价值释放的“加速器”。通过主动防御、全生命周期管理、分层架构设计和纵深防御，企业能够在模型能力提升与应用拓展的同时，将风险转化为创新动力。安全建设没有终点，唯有持续演进与制度化治理，才能保障 AI 系统可信可控，使 AI 成为稳健推动产业变革的可靠引擎。

达于行

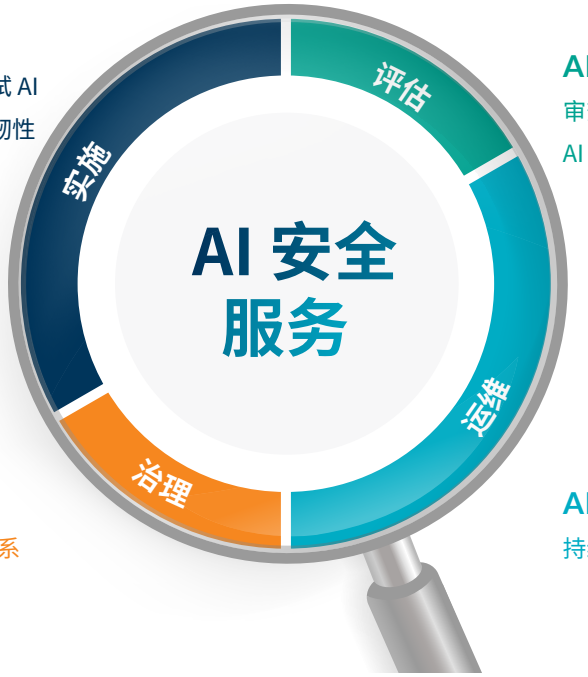
甫瀚提供的服务

我们协助组织应对当前 AI 面临的安全挑战，并探索持续改进的机遇。我们方法论的核心是 “以 AI 的速度构建安全” （Security at the speed of AI），旨在助力您提升效率与生产力、拥抱变革，并为整个组织创造 AI 领域的价值。

AI 红队演练
模拟攻击与对抗行为，测试 AI 系统的鲁棒性、安全性与韧性

AI SSDLC
识别 AI 全生命周期中的安全漏洞，确保模型从开发到部署的全生命周期安全

AI 治理
针对 AI 系统的整体治理体系



AI 安全评估
审查 AI 应用与 AI 智能体以识别风险

影子 AI 治理
提供一站式影子 AI 治理服务

AI 合规监控
持续审查模型以确保监管合规

关于甫瀚咨询

甫瀚咨询（上海）有限公司是一家具有全球视野的咨询机构。我们在中国开展业务至今已逾二十年，分别在上海、北京、深圳、成都和香港设有五个区域团队。依托甫瀚全球网络，我们能迅速汇聚甫瀚全球超过 25 个国家 90 个分支机构的资源与洞见，灵活调动更适合的专业团队为客户带来高质量的交付，并支持中国企业的海外拓展。

甫瀚咨询的业务遍及运营与财务管理绩效优化、风控与合规、内部审计、信息技术咨询、数字化转型，以及气候变化与可持续发展等领域。我们为中国各行业优秀企业、世界 500 强企业、全球各地资本市场的上市公司以及拟上市公司提供成熟及定制化的解决方案，亦为成长型企业提供陪伴式服务。

公司地址

北京
朝阳区建国门外大街 1 号
国贸写字楼 1 座 718 室
电话：(86.10) 8515 1233

上海
徐汇区虹桥路 1 号
港汇恒隆广场办公楼 1 座
2301+2310 室
电话：(86.21) 5153 6900

深圳
福田区中心四路 1 号
嘉里建设广场 1 座 1404 室
电话：(86.755) 2598 2086

成都
锦江区红星路三段 1 号
国际金融中心 1 号
办公楼 25 楼

香港
中环 干诺道中 41 号
盈置大厦 9 楼
电话：(852) 2238 0499

protiviti 甫瀚

关注甫瀚咨询
获取更多资讯