# 《欧盟 AI 法案》合规解读：机遇与挑战并存

## 敏于知

随着人工智能（AI）技术的蓬勃发展，欧洲委员会早在 2021 年 4 月就提出了《欧盟 AI 法案》（EU AI Act）草案，并于 2024 年 7 月 12 日在官方公报上公布，在公布后 20 天（即 2024 年 8 月 1 日）正式生效。该法案的目的在于加强对 AI 系统风险的识别与管控，引导企业在可持续与合规的框架下推进 AI 创新与落地。

### AI 发展带来的趋势与挑战

随着大数据、算力和算法的不断提升，AI 在各行各业得到了更深层次的应用，带来了诸多机遇，也带来了相应的挑战：

**公平性（Fairness）：**

**典型风险：**算法偏见导致招聘系统对特定群体不公，金融评分模型对少数族群或特定地区人群差异化处理，医疗资源分配算法偏向特定人群；

**应对策略：**在算法开发和数据使用阶段建立偏见检测与修正机制，通过多元化数据采样、数据偏见清洗以及决策结果公平性审计保障用户权益。

**透明性与可解释性（Transparency& Explainability）：**

**典型风险：**医疗 AI 诊断结果无法解释导致医生对结果不信任，金融贷款拒绝理由不透明引发投诉，算法决策过程对监管部门不可见导致违规；

**应对策略：**实施分层解释架构（从简单解释到技术细节），采用可解释 AI 技术（如 LIME、SHAP）。建立算法透明度分级披露制度，提供用户友好的决策解释界面，保存关键决策点审计日志，确保关键环节可追溯可验证。

**责任归属（Accountability）：**

**典型风险：**AI 决策造成损害时责任归属不明确，跨境 AI 应用面临多重法律框架挑战，算法黑箱导致责任追溯困难；

**应对策略：**技术上建立端到端决策记录机制，实施版本控制和责任追溯系统。管理上明确人机责任分工，建立 AI 决策责任矩阵，完善组织内部问责机制。

**网络安全（Security）：**

**典型风险：**AI 可能会面临网络攻击、数据泄露等安全威胁，例如模型被窃取威胁知识产权，API 接口滥用造成服务中断；

**应对策略：**技术上实施对抗训练增强模型鲁棒性，部署异常检测系统监控输入，应用加密技术保护模型参数。管理上建立 AI 模型供应链安全审查以及模型部署访问控制策略，并发展 AI 安全专家团队。

**行为安全（Safety）：**

**典型风险：**自动驾驶在罕见场景下的不安全决策，对话系统生成有害内容，推荐系统形成极化信息茧房；

**应对策略：**技术上实施行为约束和安全监控，设计人机协作控制机制，开发内容安全过滤系统。管理上建立人工审核机制，设定明确的安全边界和失效保护机制，制定应急接管流程。

**隐私保护（Privacy）：**

**典型风险：**模型记忆训练数据中的敏感信息，联邦学习中的隐私泄露，用户数据未经同意被用于训练；

**应对策略：**应用差分隐私、联邦学习、安全多方计算等隐私增强技术，实施数据最小化原则。管理上建立数据使用同意流程，实施隐私影响评估，定期进行隐私合规审计。

AI 技术的创新和应用深度不断攀升，与之配套的各类风险与挑战也日益凸显。如何在迎接 AI 新机遇的同时，做好合规和风险管理，已成为每家相关企业的当务之急。

## 《欧盟 AI 法案》的出台背景、意义与立法历程

在全球范围内，欧盟一向以严格的监管与领先的数字政策著称。欧盟之所以着手推动 AI 相关立法，正是为了在促进创新的同时，确保潜在风险得到有效管理与监管。欧盟希望通过立法将 AI 应用带来的社会与经济效益最大化，同时将数据隐私、用户安全等核心权益置于高度优先地位。

《欧盟 AI 法案》是全球首个基于风险分级体系的综合性 AI 法律框架，引领其他国家和地区的立法与监管实践，也将为企业带来明确的 AI 合规指引。

下图简要呈现了欧盟 AI 法案的相关背景与主要立法进程，该法案始于欧洲理事会对 AI 问题的讨论，并逐步推进，至 2021 年 4 月公布了《人工智能法案》提案，后续欧洲议会和欧盟成员国多次讨论修订，最终于 2024 年 5 月通过并在 2024 年 8 月起生效，该法案所包含的不同的规则将于 2025 年 2 月至 2027 年 8 月间的不同时间点生效。至 2027 年，法律将进一步完善，确保高风险 AI 应用的安全与合规性，以保护公众权益并促进技术发展。

## 《欧盟 AI 法案》的主要内容

《欧盟 AI 法案》共分为 13 个章节 115 个条款，其按不同章节（Chapter）系统阐述了 AI 应用的合规要求和监管机制，涵盖定义界定、风险分级、强制性认证及追责条款。法案基于四层风险分级框架（不可接受风险、高风险、有限风险、低风险），对 AI 系统开发、部署到退役的全生命周期实施分级监管，适用于向欧盟市场提供 AI 系统或服务的提供者、部署者及进口商等。法案要求高风险 AI 系统建立质量管控体系、技术文档披露机制及人工监督流程，对生物识别、关键基础设施等八大领域实施强制性合规认证。

此外，法案特别针对通用 AI 模型（GPAI）增设双重义务，基础通用 AI 模型提供者需公开技术文档与训练数据版权声明，算力超过 1025 FLOPs 的尖端模型需进行系统性风险评估，并对存在系统性风险的模型实施红队测试等，并设置最高达全球营业额 7% 的阶梯式处罚标准，构建起严格的 AI 治理框架。


### 定义与适用主体

AI 系统的定义广泛，涵盖机器学习、知识推理、统计学及其他数据分析型系统。只要企业在欧盟境内提供或使用 AI 服务，即在法案的管辖范围内。具体定义如下：

- **AI 系统定义：**是一种基于机器的系统，设计为以不同程度的自主性运行，在部署后可能表现出适应性，并且为了明确或隐含的目标，从其接收的输入中推断如何生成可影响物理或虚拟环境的输出，如预测、内容、推荐或决策。

欧盟《AI 法案》的适用主体采用 " 全链条 + 域外管辖 " 双维度覆盖：

- **全链条主体：**涵盖通用 AI 模型提供者及其授权代表，以及 AI 系统提供者、部署者、进口商、经销商、产品制造商、境外实体的欧盟授权代表以及其他欧盟境内受到影响的人员；
- **域外管辖：**
    - » 在欧盟市场投放使用的 AI 系统（无论实体是否在欧盟境内，含免费应用商店分发、跨境网站下载等场景）；

» 虽未直接投放系统，但处理欧盟境内数据并输出结果（如境外云服务商分析欧盟用户行为数据并将输出结果交付欧盟内运营商）；

» 开源模型开发者及算力超过 1025 FLOPs 的通用 AI 模型运营商，且将其模型作为高风险 AI 系统投放市场或投入使用。

各主体具体定义如下：

适用主体：在欧盟境内将人工智能系统（AI system）投放市场或投入使用的所有人工智能系统的运营商（operator），包括：

| 提供者 Providers | 部署者 Deployers | 进口商 Importers | 经销商 Distributors | 产品制造商 Product manufacturers | 提供商授权代表 Authorized representatives | 其他受影响者 Others |
|---|---|---|---|---|---|---|
| 指开发人工智能系统或通用人工智能模型，或让人工智能系统或通用人工智能模型得到开发并以自己的名义或商标投放市场或投入使用（无论是否收费）的自然人或法人、公共机构、机构或其他团体。（无论是否在欧盟境内） | 指在其权限下使用人工智能系统的自然人或法人、公共机构、代理机构或其他机构，但在个人非职业活动中使用人工智能系统的情况除外。 | 指位于或设立在欧盟境内，将带有第三国境内设立的自然人或法人名称或商标的人工智能系统投放市场的自然人或法人。 | 指供应链中除提供者或进口商以外，在欧盟市场上提供人工智能系统的自然人或法人。 | 产品制造商以自己的名称或商标将人工智能系统与其产品一起投放市场或投入使用。 | 指位于或设立在欧盟的自然人或法人，其已收到并接受人工智能系统或通用人工智能模型提供者的书面授权，代表其履行和执行本法规规定的义务和程序。（提供者通常不在欧盟内设立） | 欧盟境内可能受到人工智能系统影响或以其他方式受到其影响的人员。 |

## 适用范围与风险等级划分

### 通用 AI 模型（GPAI）

GPAI 直接受到 AI 人工智能法案的监管，其通常基于海量的数据训练，且具有显著的通用性，可以用于执行广泛的任务。例如大语言模型（LLM）。其通常需要添加用户接口或其他组件才能成为实际可使用的“AI系统”。

系统性风险
通用模型合规义务
+
特殊合规义务

非系统性风险
通用模型合规义务


### AI 系统

AI 系统即能够基于输入进行推理的系统，通常以软件形式出现，利用或连接多个 AI 模型来处理输入，具备一定程度的自主性，并在运行中根据实际情况进行调整，以向用户或环境输出预测、建议或决策。例如将 LLM 集成到一个 chatbot 中形成面向终端用户的完整 AI 系统。

不可接受风险
禁止在欧盟市场投放

高风险
高风险AI系统合规义务

最小风险
无合规义务

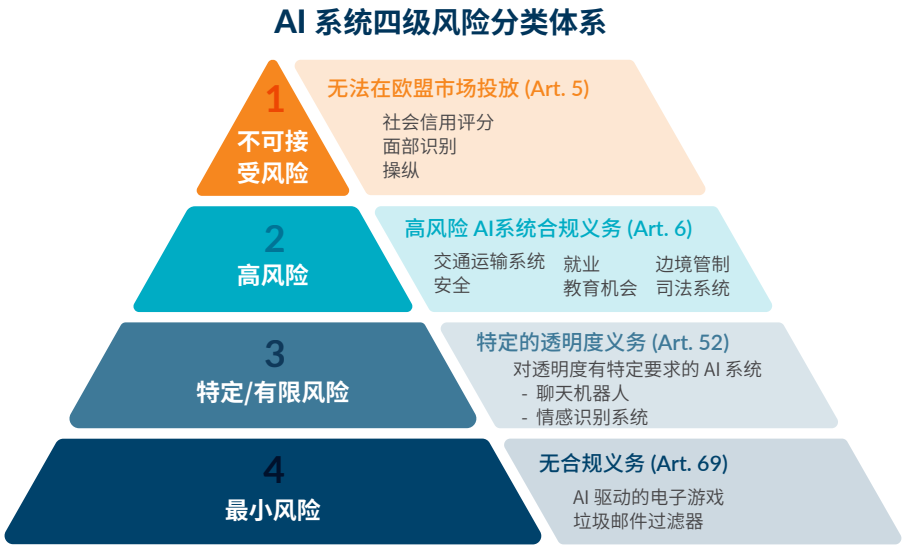
特定/有限风险
特定透明度合规义务

法案约束范围涵盖 AI 系统与通用 AI 模型（GPAI），采用风险分级与系统性风险评估结合的框架。

**AI 系统分级：**分为不可接受风险（禁止使用，如社交评分）、高风险（强制合规，如医疗 AI）、有限风险（透明性披露，如聊天机器人）和最小风险（合规豁免），不同等级对应差异化的技术与安全要求。

**通用 AI 模型（GPAI）：**除了技术透明度等通用模型义务外，还引入了系统性风险管控，评估高算力场景（如 1025 FLOPs 以上模型），要求红队测试、偏见检测、严重事件上报等特殊义务，并确保数据溯源与技术透明化，提高全球开发者的合规标准。

## AI 系统




- **不可接受风险 AI 系统**

  对于被禁止投放的 AI 系统分类，目前已于 2025 年 2 月 2 日开始实施生效。欧盟委员会保留根据技术发展与社会影响评估，通过授权法案动态调整禁止清单的权利，未来可能会禁止更多的 AI 用途。目前，无法在欧盟投放的相关 AI 系统包含以下几种类型：

| | | | |
|---|---|---|---|
| 01 潜意识操控系统 | 02 弱势群体弱点利用系统 | 03 社会信用评分系统 | 04 犯罪风险预测系统 |
| 05 系统性监控系统 | 06 实时远程生物特征识别系统 | 07 情绪识别系统 | 08 敏感属性生物分类系统 |

- **高风险 AI 系统**

  高风险 AI 系统指可能对健康、安全或基本权利造成重大风险的 AI 技术，法案中基于不同的适用主体给出了严格的合规要求。高风险 AI 系统可以分为两类：

  » 受附录一列出的欧盟线管安全立法约束或需第三方合格性评估的相关系统或产品（如医疗 AI）；

  » 附录三中规定的八大高风险 AI 应用领域（生物识别、关键基础设施、教育和职业培训、就业、基本公共服务、执法、移民和司法行政）。若仅执行辅助任务或为非关键系统则可豁免，不视为高风险场景。欧盟委员会被授权通过代表性法案来修改高风险 AI 系统清单，可以将新的 AI 系统类型添加到附录 III 中。

  高风险 AI 系统提供者需要确保其提供的系统符合七项基准，包括搭建全生命周期风险管理体系、进行数据治理以确保训练数据质量、维护实时更新的技术文档、自动记录系统操作日志、保障输出透明度、允许人类有效干预监督、维持系统准确性及网络安全。此外，法案还对高风险 AI 系统产业链中的其他主体施加了不同的合规义务：

**（1） 提供者**

**合规管理：**建立质量管理体系，实施风险评估，进行符合性评估，起草一份机器可读的 EU 符合性声明，加贴 CE 标识并在欧盟数据库注册；

**文档维护：**技术文档及操作日志至少保存 10 年；

**透明度与监督：**向部署者及其他相关方提供系统性能说明，设计可解释的 AI 输出机制；

**动态响应：**发现系统缺陷时立即召回并通知相关方，配合监管调查，一旦意识到引发严重事件（serious incidents），需要在 15 天内上报当地市场监管机构；

**（2） 部署者**

**操作规范：**指派专业人员监督 AI 运行，按系统随附的说明使用系统，并保留其控制范围内自动生成的日志至少 6 个月；

**风险防控：**实时监控系统性能，发现威胁立即暂停使用并上报；

**权益保障：**使用前进行基本权利影响评估，通知受影响的相关方；

**（3） 进口商 / 分销商**

**准入审查：**验证 CE 标识及技术文件真实性，确保供应链合规；

**风险处置：**对问题产品采取下架、召回措施，通知相关方并向监管机构通报；

**（4） 授权代表**

**文件管理及确认：**保存技术文档及合规声明至少 10 年，确保欧盟数据库注册信息准确，核实欧盟符合性声明、技术文件、一致性评估报告是否存在；

**监管对接：**作为法律实体代表配合调查，提供审计所需材料；

法案还设置了一套全链条问责体系的监管机制，通过强制日志记录（至少 6 个月）、欧盟数据库注册、沙盒外测试监管及违规分级处罚（最高罚款可达全球营收的 7%），确保风险可追溯、责任可落实。

- **有限风险 AI 系统**

有限风险 AI 系统指可能引发轻微误导但无重大危害的系统（如聊天机器人、推荐系统、文字或图像生成 AI 等），须履行透明度义务：

» AI 系统提供者需明示 AI 交互属性，确保用户与 AI 互动时获知其 AI 属性（如对话机器人标识、标注“AI 对话”等），且确保 AI 系统输出以机器可读格式标记并可检测为人工智能生成或者操作

» AI 系统部署者使用情绪识别或公共利益内容生成 AI 时（如生物特征情绪分析），必须披露系统功能及数据用途并确保输出内容标注 AI 来源。豁免情形仅适用于科研或非公开场景

- **最小风险 AI 系统**

其他未归类为不可接受风险、高风险、有限风险的 AI 系统，可被视为最小风险 AI 系统。例如，垃圾邮件过滤器、AI 视频游戏、文字处理软件、媒体平台的内容推荐算法等应用场景。

## 通用 AI 模型

### 通用 AI 模型风险分类

| 系统性风险 | 非系统性风险 |
| --- | --- |
| 通用模型合规义务<br>+<br>特殊合规义务 | 通用模型合规义务 |

此外，法案还特别针对通用 AI 模型（GPAI）增设了双重约束，所有通用 AI 模型的提供者及其授权代表，无论通用 AI 模型的部署场景或系统是否属于不可接受风险、高风险或其他风险类别，均需履行通用模型基础合规义务，具有系统性风险的通用 AI 模型则需要进一步履行特殊合规义务：

- **基础合规义务：**
  - » **技术透明度：**所有 GPAI 提供者须公开技术文档（含训练方法、数据集摘要、测试结论）及版权合规政策；
  - » **协作交付：**向下游集成商提供能力说明文档，揭露模型局限性；
  - » **长周期存证：**授权代表需保存技术文件至少 10 年并配合监管稽查；
- **系统性风险特殊义务：**触发条件为模型训练算力 ≥1025 FLOPs，或委员会判定其具有“高影响力”（如医疗决策、司法评估等场景）。相关主体须：
  - » **风险防控与模型评估：**必须使用标准化工具进行对抗测试（如红队测试），覆盖开发、上市及使用阶段，识别系统性漏洞与风险；
  - » **事件响应与透明报告：**记录所有严重事件（如公共安全威胁、重大权利侵害），并在发现后无延迟报告欧盟 AI 办公室及成员国监管机构；
  - » **网络安全强化：**确保模型及物理设施（服务器 / 存储）的端到端防护，包括访问控制、数据加密、模型防泄露等技术措施；

## 惩罚措施

根据《欧盟人工智能法案》，违规罚款金额最高可达涉事主体全球年营业额 7% 或 3500 万欧元（取高值）的罚款，累犯加重处罚。具体分级如下：

**提供禁止类AI系统**

罚款数额

前一年度全球总营业额的**7%**
或**3500 万欧元**（取高值）

罚款依据

违反第5条提供被禁止的 AI（不可接受风险的 AI 系统）

**相关责任主体违规**

罚款数额

前一年度全球总营业额的**3%**
或**1500 万欧元**（取高值）

罚款依据

相关责任主体不遵守相关规定义务，如供应商和部署者违反第50条透明度义务，供应者违反第16条，授权代表违反第22条，进口商违反第23条，分销商违反第24条，部署者违反第26条，通用AI模型提供者违反第91条或93条等

**提供不实信息**

罚款数额

前一年度全球总营业额的**1%**
或**750 万欧元**（取高值）

罚款依据

相关责任主体在通知或调查过程中提供虚假或误导信息

**例外情况**

对于中小企业，包括初创企业，以上罚款数额应以百分比或固定金额的较低者为准

# 受影响行业及合规时间表

根据欧盟《AI 法案》第 113 条的过渡期规定，不同风险等级的 AI 系统需在法案生效后 6 至 36 个月内完成合规整改。

## 重点影响行业

从事 AI 模型研发或提供 AI 系统的技术公司

为他人提供 AI 咨询、技术支持或部署服务的合作方

在产品或服务中使用高风险 AI 系统的金融、医疗、交通、教育组织，其涉及到的高风险场景示例如下：

- 金融与保险业：个人信用评分、自然人健康保险风险评估和定价
- 医疗健康业：紧急医疗患者分诊系统、AI 辅助诊断
- 人力资源与教育：简历招聘筛选系统、入学准入评估系统、监考系统
- 公共安全与执法：生物识别监控、犯罪预测系统

## 合规时间表

当欧盟 AI 法案正式生效后，将设置数个月到数年的过渡期，企业应尽早启动内部风险评估与合规整改，完成合规准备，以免影响业务正常运行。

**2025年2月2日**

受影响企业类型

高风险AI系统供应商及部署者

关键要求

禁止特定AI系统（如社会信用评分、情绪识别工具等），并开始实施AI素养强制要求（例如对医疗AI系统操作员实施年度伦理培训，并保留培训记录等）。

**2026年8月2日**

受影响企业类型

高风险AI系统运营者

关键要求

法案大部分条款生效（除第6(1)条），高风险系统需符合新规（如数据质量、透明性要求、技术文档要求等）。若高风险AI系统在2026年8月之前已上市或投入使用且在2026年8月后发生重大设计变更，需立即合规。

**2025年2月2日**

受影响企业类型

GPAI模型运营商

关键要求

2025年8月前投放市场的模型必须完成合规改造（如透明度、风险评估要求）。

受影响企业类型

高风险AI系统供应商

关键要求

第6(1)条（高风险系统分类标准）生效，需调整产品分类和合规措施。

**2030年12月31日**

受影响企业类型

大规模IT系统供应商

关键要求

在2027年8月前上市或投入使用的大型IT系统组成部分的AI系统需在2030年12月31日前完成合规要求。

## 受影响企业如何应对 — 策略与建议

在欧盟 AI 法案的监管框架下，企业尤其需要加强内控与合规策略，以确保在 AI 的应用与落地上符合欧盟法规要求。以下是我们基于实务经验与政策解读提出的专业建议：

### 评估 AI 风险，建立分级合规框架

- **风险分类评估：**根据法案标准，梳理现有 AI 系统并识别禁止的应用，优先处理高风险系统
- **通用模型评估：**评估是否开发或使用通用 AI 模型，特别关注具有系统性风险的模型
- **权利影响评估：**高风险 AI 系统部署前进行基本权利影响评估，识别对个人和群体权利的潜在风险
- **合规清单管理：**建立 AI 系统分级清单，制定对应合规计划

### 构建全面数据治理与 AI 安全框架

- **数据治理：**实施训练数据全生命周期管理，确保数据集的代表性和无偏见性，且确保与 GDPR AI 系统相关的原则保持一致（公平性、透明性、数据最小化原则等）
- **安全防护：**部署针对模型投毒和对抗样本的防御措施，确保 AI 系统的稳健性，防止泄露隐私或敏感信息
- **透明可解释：**开发分层解释工具，使 AI 决策过程可理解可追溯，并定期对透明度进行评估
- **内容标记：**为生成式 AI 内容实施机器可读的水印或元数据标记，使其可被识别为 AI 生成

### 完善内部 AI 治理，建立质量管理与监督体系

- **组织架构：**设立 AI 合规官及跨部门治理委员会，明确各层级合规责任
- **文档管理：**高风险 AI 系统提供者应建立完整技术文档体系，根据法案内容进行留存
- **监督机制：**高风险 AI 系统提供者应设计有效的人类监督措施，包括异常检测和人工干预能力
- **质量体系：**高风险 AI 系统提供者应建立质量管理系统，涵盖设计、开发到测试的全流程

### 建立持续监测与责任机制

- **监控系统：**高风险 AI 系统的提供者需要建立上市后监控系统，用以跟踪 AI 系统实际性能和偏差
- **事件报告：**制定严重事件上报流程，高风险 AI 系统提供者应确保在 15 个工作日内通知相关机构
- **供应链管理：**建立 AI 价值链各参与者（进口商、分销商、部署者等）的责任分配、合同约束以及监管报告机制，明确各方责任
- **合规更新：**建立法规变化监测机制，定期评估治理有效性并更新合规措施

## 甫瀚可提供的服务

面对欧盟 AI 法案带来的新挑战和新机遇，我们可为企业提供以下四类专业服务，助您高效完成 AI 治理与合规落地。

### AI 相关服务

#### AI 治理咨询

- 关注企业 AI 治理体系的顶层设计，包含治理组织、职责、框架、流程及路径规划

#### AI 安全及合规咨询

- 基于企业不同的 AI 部署模式及法规要求，评估相应安全管控措施的有效性

#### AI 产品开发咨询

- 探索适合于企业的 LLM 使用和应用场景，支持相关算法生命周期管理

#### AI 审计

- 关注 AI 应用的全生命周期风险及控制有效性，评估影响并提供优化建议

## 结束语

《欧盟 AI 法案》为 AI 技术发展确立了明确的监管框架与合规标准。企业需要以风险导向思维理解法案要求，将合规建设转化为提升治理水平的契机。

我们将基于专业洞见，从风险评估、制度建设到 AI 安全管控措施评估等维度提供务实支持，助力企业在合规基础上推进 AI 技术的可持续发展。

如需更多信息或定制化解决方案，欢迎与我们联系。让我们携手为企业构建一个安全、负责任与可持续增长的 AI 生态，助力企业在全球范围内获得成功。

## 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一座
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