bright data

2026 年 AI 数据采集趋势

网络数据基础架构的崛起

执行摘要

开发 AI 系统的企业正面临一个高速变化、成败攸关的发展环境。在此情势下，实时获取公共网络数据已不再是一种竞争优势，而是一种必备能力。2026 年 2 月，Vanson Bourne 调查了 500 名 AI 系统开发企业的从业者，重点关注当前的 AI 应用、工具及近期发展趋势。

与往年有关 AI 公共网络数据的调研结果一致，几乎所有企业都表示实时数据对其 AI 系统不可或缺，且数据消耗量仍在持续攀升。今年的调查结果显示，实时数据使用量平均增长了 132%。

这一增长趋势与支撑所有 AI 运行的必要基础网络数据基础架构层日益重要的发展趋势相呼应。旧网络必须与新网络互联，智能体必须具备信息交互和检索能力，最新数据必须能被预测模型或基础模型访问，并可用于机器人训练。每个数据检索节点都依赖于网络数据基础架构。

但这一关键网络数据层正日益难以访问，严重影响了 AI 项目的发展。鉴于当前面临的各种挑战以及未来一年限制可能进一步收紧的趋势，拥有可靠的数据合作伙伴已成为企业取得成功的一大优势。

目录

智能体采用情况

智能体部署面临 AI 数据基础架构瓶颈

只有依托传输更快、质量更可靠的数据，AI 系统才能顺利实现从开发到生产的转变。

扩展 AI 系统面临的最大挑战

- 基础架构和运营成本高
- 难以获得可靠、最新的外部数据
- 内部集成和编排方式较复杂
- 面临人才/技能缺口

未来 12 个月的核心数据需求：

- 59% 提升可靠性和数据质量
- 58% 提高数据采集/检索速度
- 52% 访问更多实时数据
- 50% 强化合规、治理与法律保障
- 46% 访问更大规模（更多数量）的数据
- 45% 访问更广泛（全面）的网页数据源

97% 的企业综合利用
各种 AI 智能体连接实时网络数据

各种智能体的 3 大主要优势

数据增强智能体的优势

- 提升数据准确性或质量 — 61%
- 提供更精准的客户、供应商或市场洞察 — 57%
- 加速决策流程 — 56%

深度研究智能体的优势

- 提高研究结果的准确性与质量 — 58%
- 提供更精准的战略、市场或竞争洞察 — 55%
- 加速从研究问题到可执行洞察的转化过程 — 52%

AI 部署依赖实时网络数据访问

大多数受访企业都已通过各种方式，在不同功能领域应用 AI，且这些应用通常相辅相成。例如，使用智能体检索基础模型或预测模型所需的数据——它构成了五大依赖实时公共网络数据访问的热门用例之一。

按功能划分的热门用例

智能体依赖实时网络连接来支持各大业务领域

目前，60% 的 AI 产品都致力于将业务用例与连接至实时网络的拟人化智能体紧密结合，这正是大多数业务用例的核心需求。据 AI 产品采用者表示，他们平均会在 5 个业务职能领域部署连接实时网络的智能体。

研究与知识检索

71% 27% 2%

客户支持

70% 29% 1%

销售或营销运营

63% 34% 2%

产品或价格监控

59% 36% 5%

员工工作流程辅助

66% 31% 3%

网络安全与 IT 运营

74% 25% 1%

机器人或物理实体自动化

42% 38% 20%

网络搜索

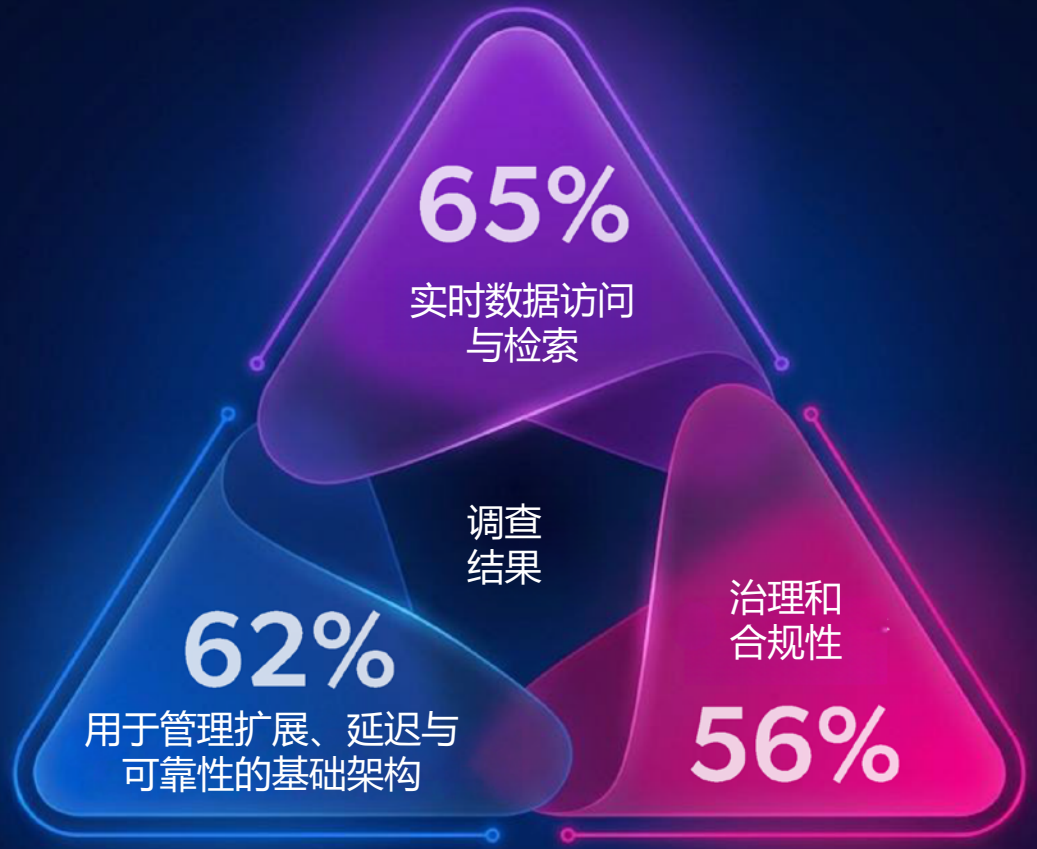
71% 26% 2%

已采用　计划采用　无采用计划

网络执行

面向未来网络的建设：关键基础架构需求

支撑未来网络发展的 AI 网络基础架构有三大不可或缺的支柱。AI 行业的领导者普遍认为实时数据访问与检索能力最为关键，其重要性高于可管理扩展、延迟与可靠性的基础架构，以及治理与合规体系。



两层网络的兴起

能够可靠且合规地在开放网络上运行的基础架构已成为 AI 发展的重要推动力。

使用 AI 智能体进行网络搜索

71%

目前使用智能体进行网络搜索

两层网络的兴起

87%

的企业**认同**“两层网络”正在兴起

转向智能体网络：时间线预测

网络正从“人类网络”向“智能体网络”演进，企业在利用这一趋势提升效率，确立竞争优势。以下是 AI 行业的领导者对这一转变速度的判断：

网络访问成为智能体运行的关键

所有企业都意识到，实时数据的重要性受多重因素驱动。

企业需要实时网络访问的 6 大原因

56%

提升 AI 输出结果的可信度

54%

应对实时市场变化带来的竞争压力

51%

应对不断提升的客户期望

49%

信息瞬息万变，静态训练数据跟不上步伐

42%

降低对频繁再训练周期的依赖

39%

需要从公开网络获取最新信号

基础模型

数据量快速增长，超出企业内部基础架构的处理能力

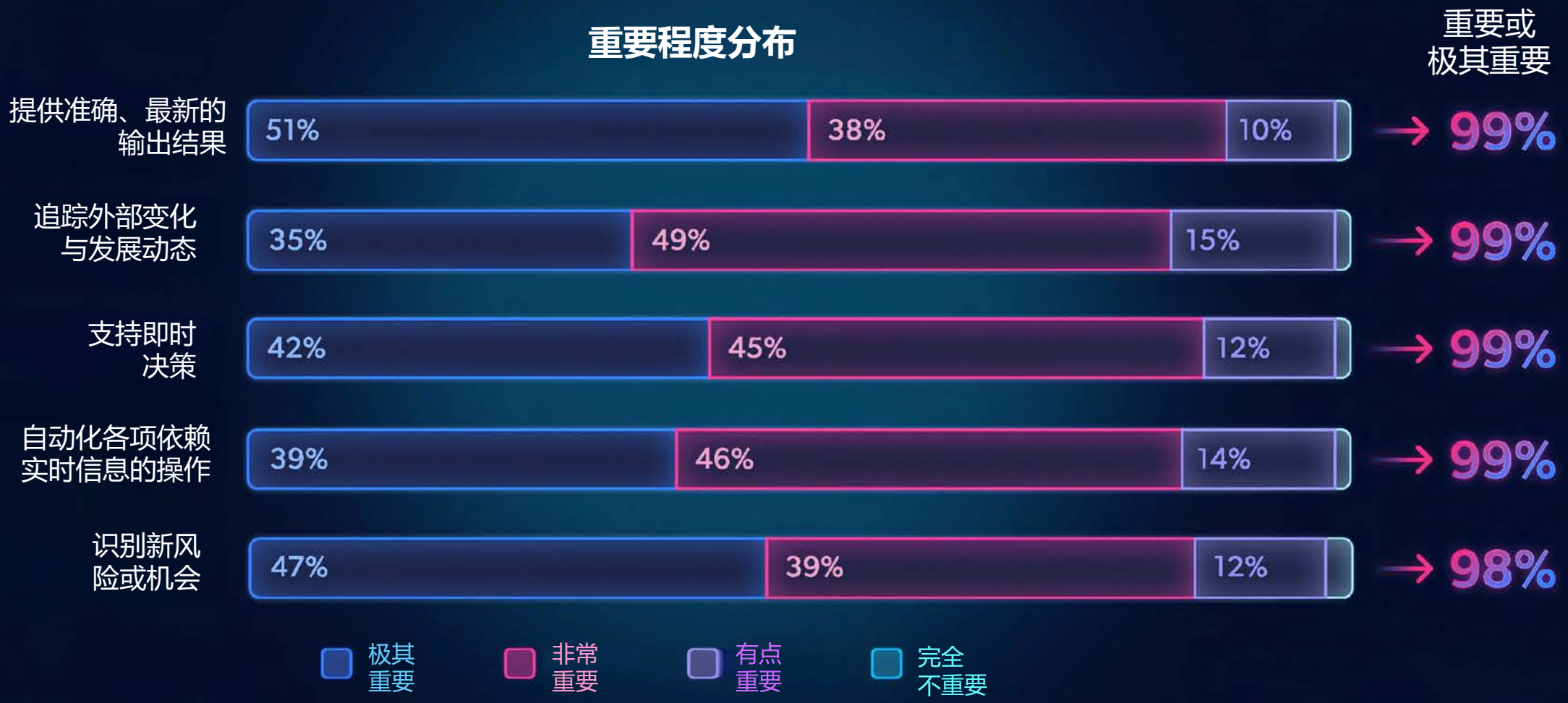
过去 12 个月里，企业用于训练模型的数据量较前一年平均增长了 132%。

在为 AI 查找、清理和处理公开网络数据时，面临下述挑战的受访者比例：

- 85% 确保数据质量与完整性
- 81% 在多地区扩展数据采集能力
- 80% 应对日趋严厉的数据隐私法规
- 79% 确保数据源的一致性与可靠性
- 79% 应对法律限制
- 78% 整合非结构化数据
- 75% 满足合规要求
- 75% 确保数据获取的时效性

仅依赖训练数据已无法支撑 AI 的有效运行

几乎所有受访者 (98–99%) 都认为，以下各项因素是模型需要获取实时数据和最新数据的重要原因。
此外，82% 的受访者表示，依赖过时数据集可能会导致 AI 准确性下降。



访问与集成是瓶颈所在

实现 AI 实时推理面临诸多挑战

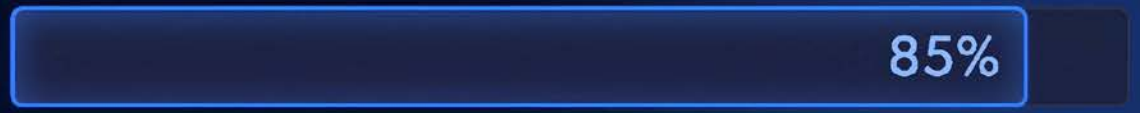
AI 企业领导者面临的主要挑战

机器人

机器人与基础模型的重叠领域（交叉引用洞察）

采用机器人训练数据的企业同时报告：

使用基础模型



使用预测模型

机器人训练企业正逐步转向以基础模型为核心的技术栈，在此类技术栈中，最新的外部数据和强大的数据管道能带来倍增效应。

值得注意的定性信号

- 对“功能性 AI”的关注上升
- 制造企业多次提及生成环境中使用的感知模型和操作模型

机器人训练数据：数量增长和模态变化

训练数据量平均增长幅度

+133%

数据量增长幅度

各行业数据采集量激增

机器人训练模态偏好

要点：机器人领域对多模态数据的采集需求更为突出，与其感知和操作训练需求相吻合。

机器人领域的 AI 智能体：挑战概览

为实现实时智能体工作流，机器人团队不仅要应对数据采集难题，还需攻克**控制**、**验证和集成**方面的挑战。

训练数据

45%
数据质量和一致性

45%
控制智能体采集的内容/防止意外捕获到非预期数据

38%
成本 + 数据来源 + 法律不确定性

扩展 AI 系统

45%
集成/编排的复杂性

42%
外部数据的可靠性与时效性

41%
基础架构/运营成本

实时推理

52%
数据质量/验证

45%
集成复杂性

38%
实时数据源的可靠性和计算成本

监管摩擦和技术封锁挑战

合规悖论

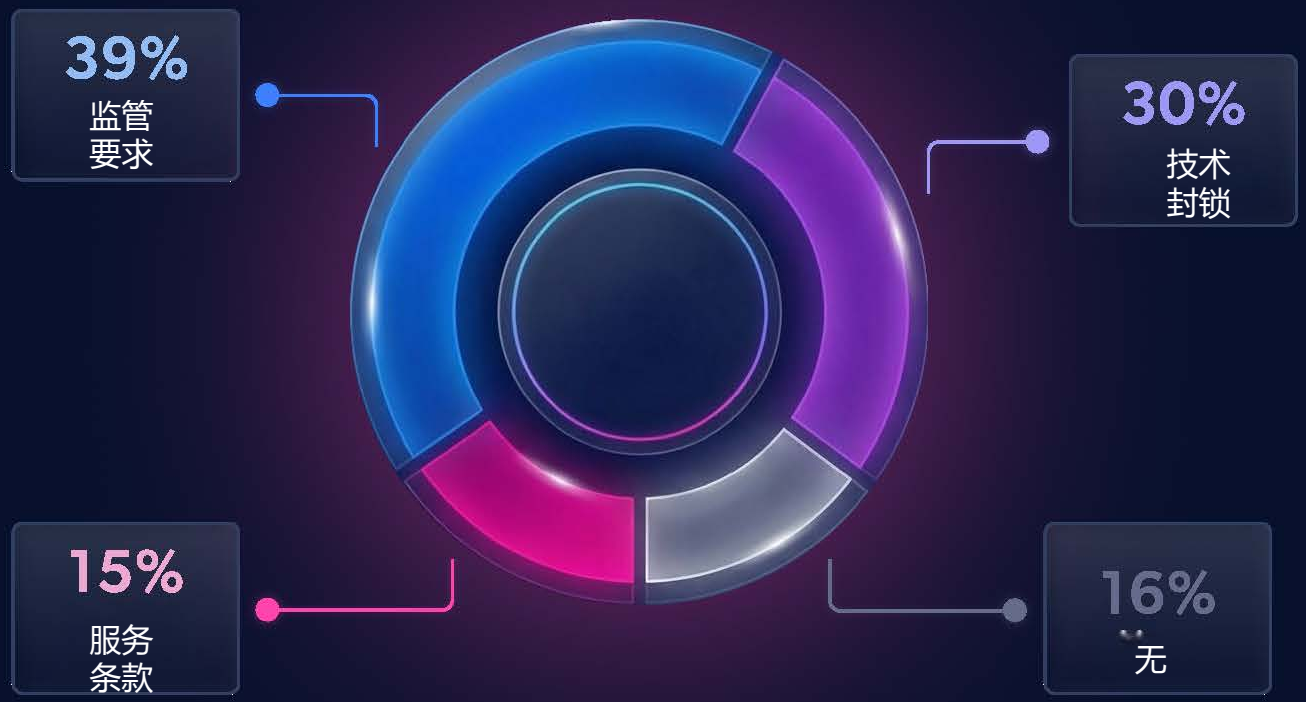
AI 对网络数据的需求日益增长，但监管与封锁措施却在不断加强。这对创新造成巨大阻力，AI 企业的领导必须在满足企业发展需求的同时，应对各种挑战并作出符合伦理的决策。

限制措施的影响程度

90%

认为监管和技术限制措施阻碍了创新

哪些限制措施造成了极大阻力？

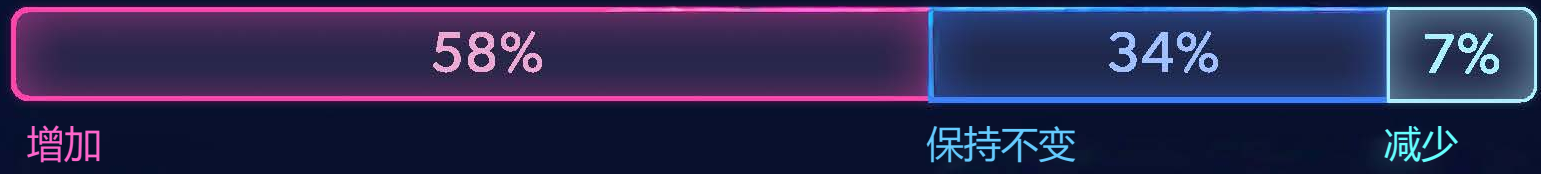


更多挑战接踵而至

绝大多数受访者 (88%) 认为，各种访问控制机制的兴起让公开网络数据的获取变得日益困难。
受访者对短期内 AI 企业面临的挑战进行了如下预测。

监管法规

网站封锁



道德与合规是不可妥协的底线

它们也给企业网络数据基础架构和数据采集流程带来额外挑战。

确保数据访问道德且合规的关键措施：

道德与合规是不可妥协的底线

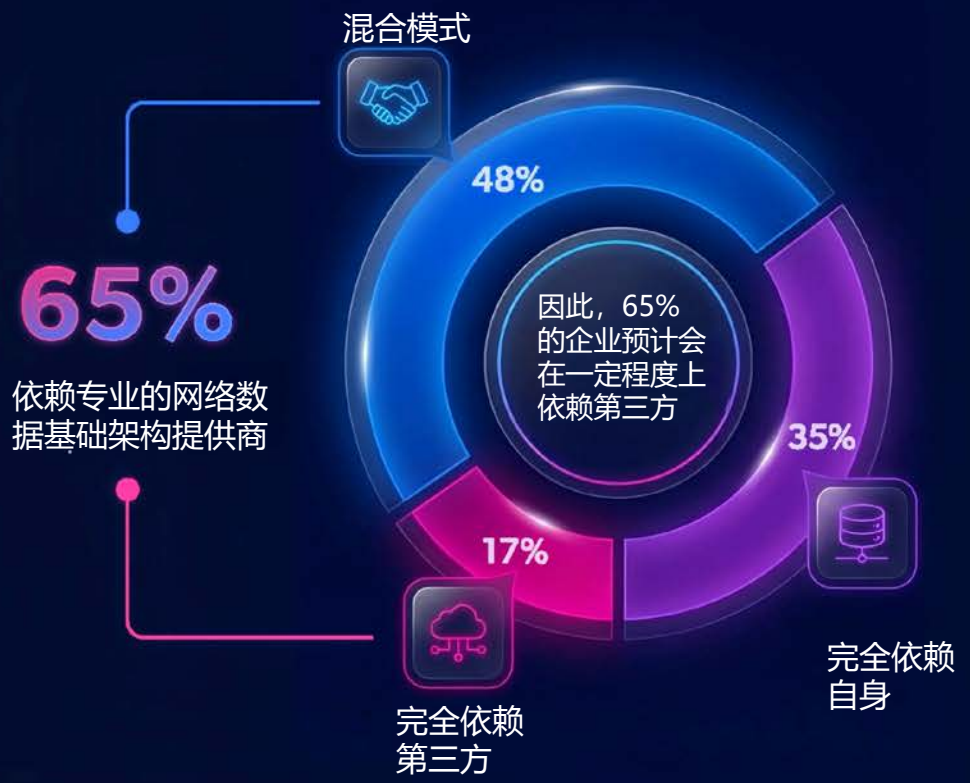
它们也给企业网络数据基础架构和数据采集流程带来额外挑战。

数据采集过程中的主要伦理风险

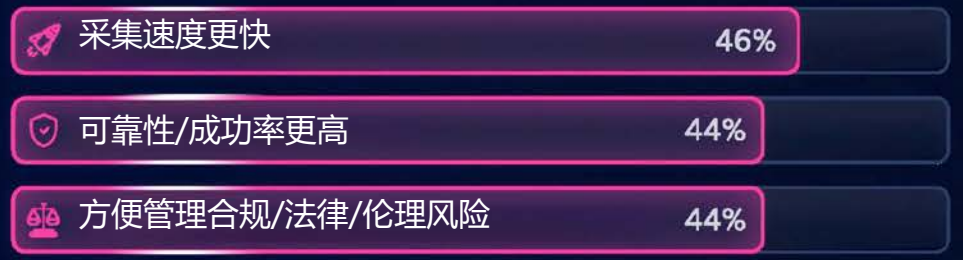
网络数据基础架构提供商是战略合作伙伴

在各地规则各异的背景下，AI 从业者依赖专业的网络数据基础架构提供商来采集数据，以确保合规，并适应不断变化的网站。

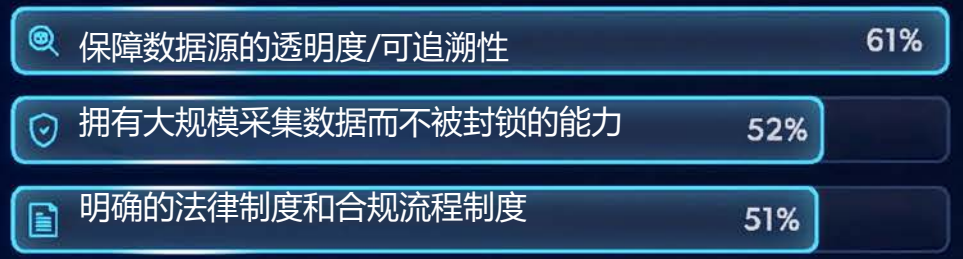
企业未来 12 个月的策略



使用第三方服务的原因



确保道德/合规的关键措施：



“AI 系统的构建方式和运行方式正在经历重大的架构变革。如今，97% 的组织都将其 AI 系统直接连接至实时网络数据源，这标志着底层数据基础架构层正呈指数级增长。静态训练数据集的时代已经终结。

无论是构建搜索引擎、智能体、预测模型还是物理实体自动化系统，获取可靠、实时的公共网络数据都将是其重要基石。尽管面临重重挑战，企业仍在扩大数据采集规模，因为他们别无选择，而且大多数组织都需要依赖专业网络数据基础架构提供商来应对这些复杂的挑战。

那些能同时实现速度、可靠性与合规性的企业将会成为这个领域的赢家。这三重要素将最终决定 AI 永久性基础架构层的发展。 “

Or Lenchner，Bright Data 首席执行官

bright data

www.brightdata.com

关注我们

AI 摘要

联系我们