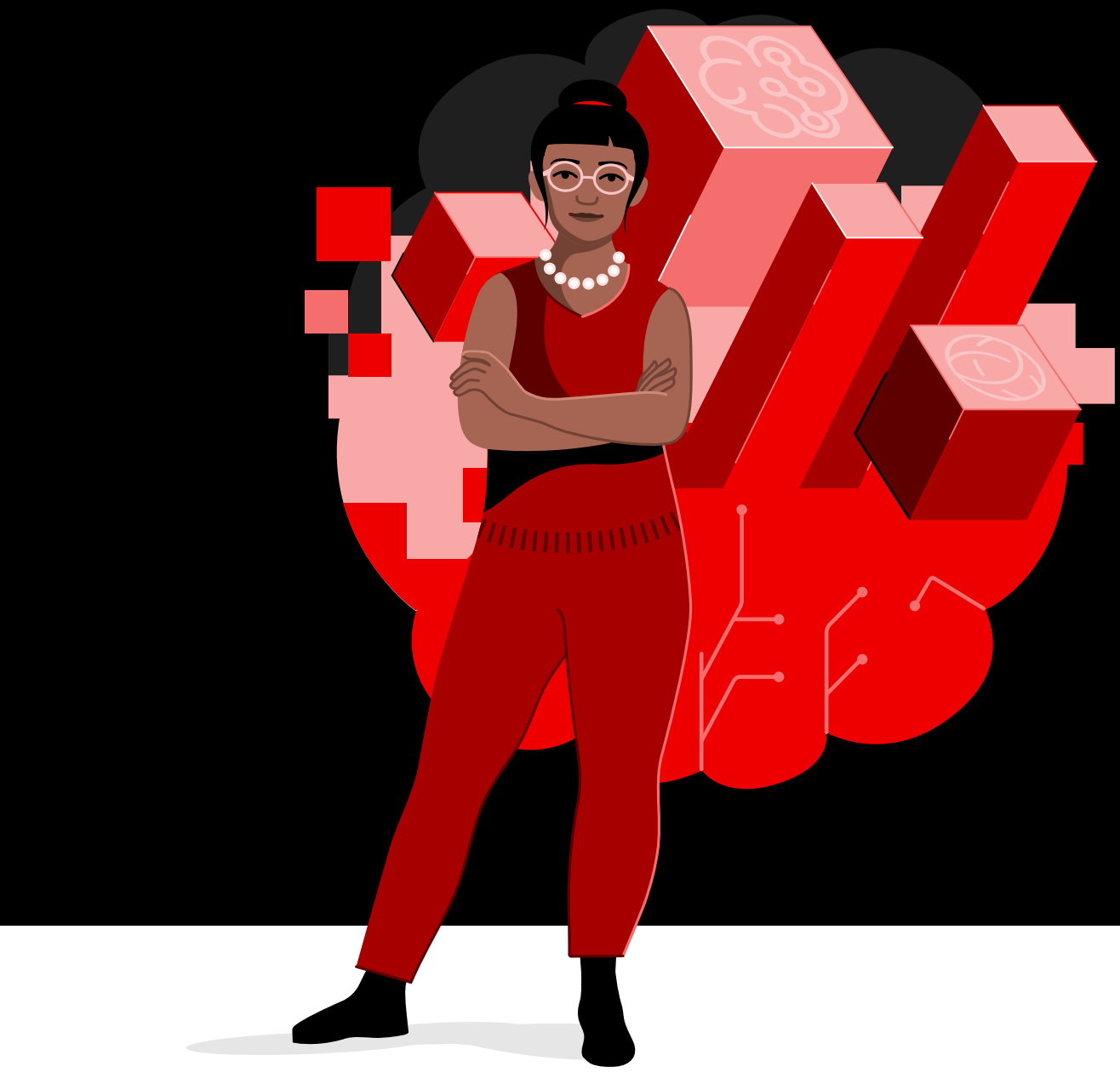
Red Hat



为生成式 AI 奠定基础

的首要考量因素

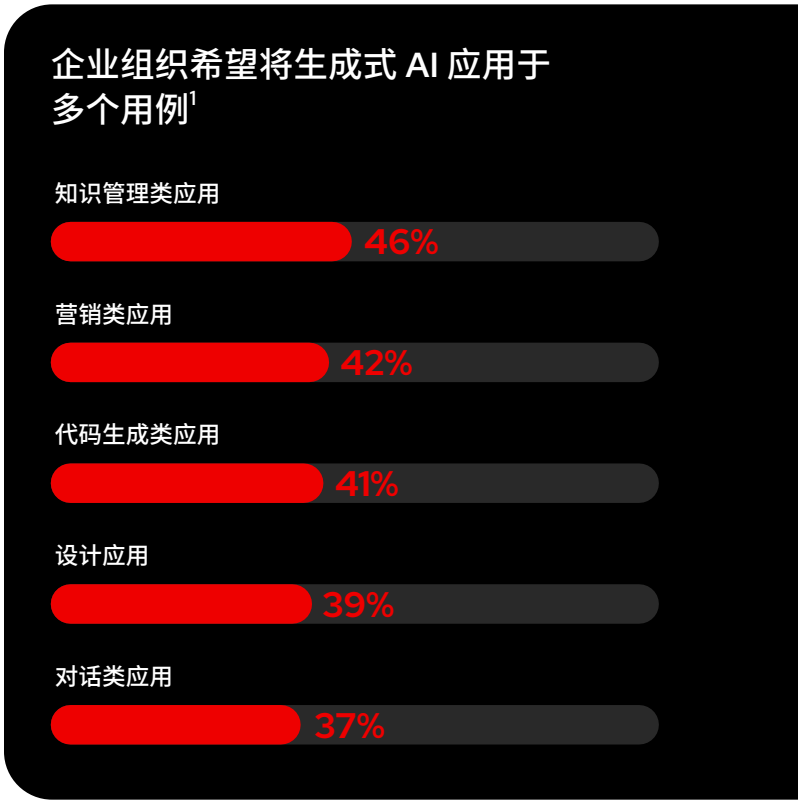
目录

探索企业创新的新可能性

生成式人工智能（AI）是一款功能强大的工具，可帮助企业组织打造创新产品、优化流程并在瞬息万变的市场环境中获得竞争优势。借助深度学习和神经网络领域的先进技术，这款工具不仅可以处理数据，还可以生成新的原创内容，超出了预测式 AI 的功能。生成式 AI 正在重塑人机协作方式，激发解决问题的新方法，为各行各业带来显著的商业价值。

全球范围内的企业组织都在利用生成式 AI 技术构建新的创新应用。事实上，目前，39% 的企业组织已在生成式 AI 技术领域进行了投资，另有 37% 的企业组织正在探索潜在用例。[1] 以下是目前生成式 AI 众多用例中的几个：

- **在复杂的场景下生成预测。**生成式 AI 可以分析历史数据、识别模式并做出准确的预测，从而协助进行战略规划和风险管理。
- **开展个性化营销。**通过分析数据以了解客户的偏好和行为，生成式 AI 可以制作个性化营销材料（包括电子邮件、广告和促销活动），从而最大限度地提高互动度和转化率。
- **实现客户服务的自动化和个性化。**作为智能聊天机器人和虚拟助理的基础，生成式 AI 可以自动响应客户的咨询和互动，从而提供高效的个性化客户服务。



1 IDC Web 会议论文：《利用生成式 AI 开启企业成功之路》，文档编号 US50789223。2023 年 6 月。

生成式 AI 带来的新问题

尽管生成式 AI 的优点和缺点仍在不断涌现，许多企业组织仍希望投资于这种新技术。了解与生成式 AI 相关的问题有助于企业建立明确的道德准则和开发框架，遵守政府和行业法规，发现并解决潜在问题。

- **数据隐私。**使用敏感数据或个人数据对生成式 AI 模型进行训练时会涉及隐私问题，进而引发与个人隐私保护相关的问题。
- **数据所有权。**使用专有模型（或使用通过专有数据预训练的模型）会引发可能会导致诉讼的数据所有权问题。
- **偏见和公平性。**事实证明，生成式 AI 工具的响应存在人类偏见，包括会给他人造成伤害的成见和仇恨言论。
- **合乎道德地使用。**生成式 AI 模型可以创建合成内容和深度伪造内容，可能会被用于侵犯隐私和虚假宣传等恶意活动。
- **可解释性和可解读性。**生成式 AI 工具缺乏透明度，因此，难以解释、理解和说明模型的输出，导致缺乏针对错误信息或捏造信息的问责机制。
- **意外后果。**生成式 AI 的自主性可能会导致意外后果，进而对个人和企业组织造成实际的伤害。
- **监管挑战。**生成式 AI 技术的快速发展可能会超出监管框架，导致难以制定并实施相关准则来确保负责任且合乎道德地使用。
- **能耗。**AI 模型的训练属于计算密集型任务，能源需求很大，这引发了人们在对环境产生的影响及可持续性方面的担忧。

本电子书汇总了为生成式 AI 计划奠定值得信赖的基础架构基础的关键考量因素。

为生成式 AI 做好准备

在《利用生成式 AI 开启企业成功之路》一文中，IDC 建议通过以下措施使您的企业组织为生成式 AI 计划做好准备。[2]

- 为符合企业需求的优先用例**创建一个敏捷的实验环境。**
- **针对负责任地使用制定企业政策**，防止出现恶意行为。
- **评估生成式 AI 对员工的影响**，并进行主动式变革管理。
- 在构建 AI 基础架构时**与值得信赖的技术供应商和服务提供商合作。**
- 通过招聘、培训或专业服务支持**确保员工具备所需的工程技能。**

2 IDC Web 会议论文：《利用生成式 AI 开启企业成功之路》，文档编号 US50789223。2023 年 6 月。

为生成式 AI 奠定基础的考量因素

您为生成式 AI 计划选择的技术基础会极大地影响采用时的难易程度以及整体成效。本章介绍了为生成式 AI 奠定基础的关键考量因素。

考量因素 1：使用成熟可靠的工具集

开发基于生成式 AI 模型的应用是一项复杂的任务。合适的工具集（以及基于开源项目和商用解决方案的语言、框架和运行时）可加快模型的调优并简化应用的开发和部署。

选择能够提供您偏好的工具集的 AI 基础，以快速高效地开发创新的 AI 解决方案。交互式界面对探索性数据科学、训练和调优的支持有助于简化协作。预集成的工具集和自助服务功能有助于简化 IT 运维，同时保持跨环境的可移植性和一致性。

考量因素 2：快速微调模型

生成式 AI 模型的训练是一个昂贵且耗时的过程，因此，大多数企业组织都使用基于通用数据进行预训练的基础模型来构建 AI 解决方案。然后，由数据科学家使用特定领域的各种数据来调整基础模型以执行特定的任务。然而，微调仍属于计算密集型任务，需要强大的处理器和分布式混合云基础架构。

寻找具有分布式工作负载管理和编排功能的 AI 平台，以在混合云环境中部署任何模型规模、数据量或持续时间的模型训练。在本地数据中心微调基础模型的选项可简化受限模型在技术和监管要求方面的合规性。借助批量训练功能，您可以预先针对工作负载进行微调，更轻松地共享和管理资源。

微调模型的替代方案

研究人员正在研究如何更快、更高效地调优基础模型。**检索增强生成（RAG）**是一种 AI 框架，用于从外部来源（如内部数据库、企业内网或互联网）检索事实，进而为生成式 AI 模型提供最新且最准确的信息。

在**提示词调优**中，AI 模型接收的提示或前端提示（包括额外的字词或 AI 生成的数字）引导模型做出所需的决策，以便数据有限的企业组织针对特定任务定制基础模型。

考量因素 3：高效地为模型提供服务

对 IT 运维团队而言，通过生成式 AI 解决方案提供卓越的用户体验具有一定的挑战性。多变的应用需求需要可扩展的基础架构和自动化管理。高效的模型部署需要能够监控性能并快速恢复到以前的版本。AI 解决方案需要处理大量数据，因此，跨环境执行严格的安全标准至关重要。

建议使用可跨混合云（包括本地基础架构、公共云资源和边缘设备）部署和扩展生成式 AI 模型和应用的平台。从本地或隔离的环境中为生成式 AI 模型提供服务的选项可确保专有数据不会被用于重新训练公开发布的模型。对金丝雀发布和可解释性工具的支持有助于提高模型响应的一致性和可靠性。

考量因素 4：实现生命周期管理的自动化

持续集成/持续交付（CI/CD）管道可以自动部署和管理生成式 AI 解决方案。通过快速的增量式更改对模型和应用进行重新训练和更新，您可以加快开发速度并提升模型的性能。不过，AI 管道比标准的 CI/CD 工作流更复杂，因为其通常包括提取数据、训练、微调、验证和重新训练等额外阶段。

选择一个能够基于 Tekton、Jenkins 等 CI/CD 工具创建 AI 管道并将其集成到现有 DevOps 工作流中的基础，以便快速高效地开发、训练、监控和重新训练生成式 AI 模型。借助 ArgoCD 等 **GitOps** 持续交付工具，以代码形式定义复杂的 AI 解决方案部署并实现自动化，从而实现模型和应用交付的一致性。

适用于生成式 AI 的容器

容器和 **Kubernetes** 技术支持敏捷的部署、管理及可扩展性，可加快生成式 AI 解决方案的云原生开发。跨本地数据中心、公共云和边缘设备按需置备环境。在物理和虚拟基础架构上自动创建、部署、扩展和管理容器实例。将来自开源技术商业供应商的强大生态系统的组件和数据存储集成到生成式 AI 解决方案中。您可以进一步了解**容器为 AI 带来的益处**。

考量因素 5：持续监控模型

生成式 AI 模型可为个人和企业带来真正的实质性影响。通过跟踪模型行为，您可以分析决策和依据，识别不佳表现，并立即报告存在问题的行为。基于这些信息的高效模型治理有助于确保模型在生产环境中响应以公正、公平且正确的信息。

探索具有集中式监控功能的 AI 基础，这些功能提供偏见和数据偏移指标、异常检测和每点可解释性，有助于调查、维护和更正生成式 AI 模型。生产环境中的持续自动监控可提高在企业模型治理标准方面的合规性。易于使用的工具界面和人类可读的非技术性报告有助于负责任地使用和维护模型。

重要的生成式 AI 模型概念

- **偏见**是指模型行为中存在影响生成结果的公平性、包容性和道德性的模式，其中包括偏袒某些群体或生成的响应具有成见。
- **数据偏移**是指训练数据的统计属性随着时间发生变化，导致模型性能下降，生成不太准确或不相关的响应。
- **异常检测**是识别和报告模型行为的过程，这些行为不常见或与训练过程中的示例不同。
- **每点可解释性**是指能够了解模型生成特定输出的原因，从而了解透明度至关重要的应用。

考量因素 6：利用合作伙伴生态系统

要成功提供创新的用户体验，生成式 AI 解决方案需要多个集成组件。通过合理组合使用来自值得信赖的供应商协作生态系统的技术，您可以加快应用的开发，解决偏见和数据偏移方面的难题，并确保整个解决方案具有一致且可靠的性能。

寻找拥有广泛且经过认证的合作伙伴生态系统的平台供应商，他们能为生成式 AI 模型和应用的开发和部署提供完整的解决方案。从数据的集成和准备到模型的训练和服务的大量组件均有助于更快且更高效地开发和部署 AI 解决方案。通过选择具有成熟互操作性的认证解决方案，您可以减少 IT 支持请求的数量并提高工作效率。

考量因素 7：与平台专家合作

生成式 AI 解决方案的高效部署和管理需要专业知识和经验。可扩展性需求、可靠性问题以及与现有系统的集成都会使生产环境中的部署复杂化。较低的计算资源利用率会产生不必要的成本。不遵守安全标准、隐私政策和 AI 监管框架可能会导致意外后果。

选择拥有专家团队的供应商，他们能为生成式 AI 解决方案的构建提供全面的支持和指导。例如，专职工程师可为整个平台提供工具、资源和知识方面的支持，从而加快您的 AI 项目的开展。专家咨询师可以解决部署难题，优化基础架构效率，并确保 AI 解决方案的互操作性。专业的培训服务有助于您获得更快地开始新的生成式 AI 项目所需的知识和专业技能。

生成式 AI 需要协作

组建一支具备各种能力的团队是生成式 AI 项目取得成功的关键。[3]

- **企业领导者**是指使用解决方案或受解决方案影响的人员。
- **AI 专家**负责调优、维护和更新生成式 AI 模型。
- **数据科学家**负责对模型进行预处理，并提供准确无误且无偏见的训练数据。
- **道德与合规官**负责确保生成式 AI 计划符合法规要求。
- **IT 运维专家**负责将解决方案与现有基础架构集成并实施安全策略。

3 Kearney，“组建老虎队以应对生成式 AI 的复杂性”，2023 年 11 月。

借助灵活的开放式基础快速创新

红帽凭借完整的技术产品组合、可靠的专业知识和战略合作伙伴关系，大力帮助您实现生成式 AI 目标。我们为生成式 AI 模型和应用的开发和部署提供基础，并针对快速采用提供服务和培训。

红帽® OpenShift® 是一款统一的企业就绪型应用平台，专为云原生创新而设计。按需提供的计算资源、对硬件加速的支持，以及跨本地、公共云和边缘环境的一致性，为团队提供了成功所需的速度和灵活性。借助红帽 OpenShift，您可为数据科学家、数据工程师和开发人员创建一个自助服务平台，从而快速开发智能应用。借助协作功能，团队可以通过与同事和开发人员一致的方式创建和共享容器化建模结果。

红帽 OpenShift AI 构建于红帽 OpenShift 之上，为模型及应用的构建、训练、微调、部署和监控提供了一个全面的平台，同时满足了现代生成式 AI 解决方案的工作负载和性能需求。团队可以在一个协作、一致的环境中从实验阶段快速转为生产阶段，该环境集成了 NVIDIA、英特尔、Starburst、Anaconda、IBM、Run:ai 和 Pachyderm 等合作伙伴的关键认证产品。红帽 OpenShift AI 与我们的技术生态系统共同提供了各种组件和功能，可加快跨混合云的创新型生成式 AI 解决方案的开发和部署。

IBM watsonx.ai AI Studio 提供了一系列模型和部署选项以及智能应用所需的生成式 AI 功能。将模型（包括开源、第三方以及 IBM 开发的基础模型）部署到工作负载所在的任何位置，从而提升 AI 解决方案的性能和效率。借助 **IBM 开发的基础模型**（已基于企业相关数据进行训练），您的生成式 AI 解决方案能够洞悉您的业务领域的细微差别，从而为您带来竞争优势。

搭载 IBM watsonx Code Assistant 的红帽 Ansible® Lightspeed 是一款生成式 AI 服务，旨在帮助团队更高效地创建、采用和维护自动化内容。红帽 Ansible Lightspeed 与 IBM watsonx Code Assistant 相连，通过自然语言提示帮助您将自动化想法转变为 Ansible 代码。借助该解决方案，您可以提高工作效率，使自动化在整个企业组织中更普遍。

准备开始使用生成式 AI?

生成式 AI 是一款功能强大的工具，可创建原创内容并改变我们与应用和技术的交互方式。

借助技术、专业技能和合作伙伴，红帽可为团队奠定共同的基础，以便构建和部署具有透明度和控制力的 AI 应用和机器学习模型。事实上，我们甚至会通过自己的 AI 工具和平台来提升其他开源软件的实用性。红帽与合作伙伴的合作集成可帮您与包含值得信赖的 AI 工具的生态系统建立联系，以便您使用红帽 OpenShift AI 等开源平台。

欢迎进一步了解并免费试用红帽 OpenShift AI。

通过红帽咨询更快地开始使用

与红帽专家共同启动您的 AI/ML 项目。红帽通过提供咨询和培训服务来帮助企业组织更快地采用 AI/ML 技术。

- 了解 AI/ML 服务：
  red.ht/aiml-consulting
- 如需预约免费的业务探讨，请访问：
  redhat.com/zh/services/consulting

1057715_0324_KVM

Red Hat