Red Hat AI

生成式 AI 的实际应用

考量因素及红帽 AI 的实际应用

目录

生成式 AI：重塑市场和行业

人工智能（AI）仍是全球企业组织主要的创新和投资领域。事实上，据 IDC 预计，全球 AI 解决方案支出的复合年增长率（CAGR）将在 2023 至 2028 年期间达到 29.0%，增长至 6,320 亿美元[1]。

生成式 AI（gen AI）是这一增长的关键驱动因素，预计同一时间段的全球支出复合年增长率将达到 59.2%[1]。生成式 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 基于从现有信息中学习的模式来生成新的内容或数据。它可以生成与训练数据相似的文本、图片、代码、声音或其他媒体，为内容的创作和个性化提供创新解决方案。因此，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。

根据 IDC 的数据，到 2028 年，全球对生成式 AI 解决方案的投资预计将超过

2020 亿美元，

2023 年至 2028 年的复合年增长率将达到 59.2%[1]。

生成式 AI 应用可为企业组织带来诸多优势：

- 提高员工的工作效率。
- 提高客户满意度。
- 降低运维成本。

本电子书介绍了选择 AI 解决方案时的关键策略和考量因素，分析了在即用型方案与定制开发方式之间实现平衡的解决方案所具有的优势，还介绍了企业开启生成式 AI 之旅的常见用例。请继续阅读，了解如何为生成式 AI 创新奠定基础。

1 IDC，“IDC FutureScape：2025 年全球人工智能和自动化预测”，2024 年 10 月 28 日，文档编号：US51666724。

选择适合您的企业组织的 AI 战略

与任何大型 IT 或业务举措一样，为企业组织制定实施 AI 的战略对于成功至关重要。

企业在制定 AI 战略时有两条路径可以选择：采用基于云的 AI 服务；或者自行构建并托管 AI 平台。这两种方式在技术参与度和运维投入方面的要求有所不同，所提供的定制化程度与可控性也存在差异。

基于云的 AI 服务

基于云的 AI 服务是由第三方供应商提供的一种付费托管解决方案。这类服务通过应用编程接口（API）提供对前沿模型的访问权限。您的企业组织无需自行托管 AI 模型即可将其集成到您的应用中。一些私有商业解决方案还允许您对提供的模型进行微调，或者将模型部署到专用或管控更严格的环境中。

由于这种方式提供了即用型 AI 解决方案，且与模型本身的交互最少，因此，对于不想处理 AI 基础架构管理的复杂性、运维团队规模较小或在小规模上采用 AI 的企业组织来说，这是一种更简便且更具成本效益的选择。

自托管 AI 平台

自行构建并托管 AI 平台能让您在模型和环境方面拥有更多选择和控制权。您可以根据企业组织的需求，自由选择最合适的硬件、软件、模型、应用及部署位置。例如，您可以选择将模型和应用托管在公共云、私有云、本地数据中心或边缘位置。这种方式还为您提供了更多定制模型和应用的机会，让您对数据有更强的控制权，同时减少了对第三方提供商的依赖。即便如此，与基于云的 AI 服务相比，这种方式通常需要更高的前期投入以及更多的持续运维工作和维护成本。

要构建并托管 AI 平台，您需要：

- 拥有适合您的用例的基础模型。基础模型示例：大语言模型（LLM）、代码模型、小语言模型（SLM）、开源模型及多模态模型。
- 能够使用图形处理单元（GPU）等硬件加速功能。
- 能够使用具有高级 AI 工具和服务机制的应用平台。
- 用于合规管理及负责任地使用 AI 的治理解决方案。

这种方式能让您对 AI 解决方案有更强的控制权，因此，如果企业组织身处监管严格的行业、计划在 AI 解决方案中使用敏感数据和知识产权（IP）或是拥有较大规模运维团队（能够应对 AI 基础架构的构建、运行及维护等复杂工作），这种方式无疑是一种理想选择：

AI 战略实施路径对比

	基于云的 AI 服务	自托管 AI 平台
部署	+ 通过即用型解决方案实现更快的部署	– 部署速度较慢，且需要更多规划
成本	+ 前期成本较低 – 潜在的隐性成本，尤其是在规模化和定制化方面	– 前期成本较高 + 无隐性成本
数据隐私与安全性	– 数据隐私、安全性及知识产权风险较高，且控制权较弱	+ 在本地部署时可提升数据隐私与安全性
解决方案定制化	– 定制能力有限 – 供应商锁定及依赖性	+ 完全定制能力 + 对供应商的依赖程度较低
技能要求	+ 由于已包含硬件、模型和支持，因此所需技能极少	– 需要 AI 基础架构和运维技能
最适合	不打算自行管理 AI 基础架构的企业组织	希望在其 AI 解决方案中拥有更多控制权和定制化的企业组织

评估 AI 解决方案时的考量因素

评估 AI 解决方案和战略时，请务必考虑其透明度、效率和相关性。

确保 AI 解决方案中的透明度

AI 解决方案能否提供透明度、可问责性和可解释性，同时确保数据隐私、安全性和监管合规性，对于建立信任、降低风险和保持竞争力至关重要。寻找那些能够清晰披露模型架构、训练数据和性能指标且能够为 AI 生成内容提供问责机制和解释的供应商。

优化基础架构和成本效益

支持模型优化、分布式训练和高效硬件配置的可扩展、低成本基础架构解决方案能够帮助您最大限度地降低运维开支，提升性能，并快速适应不断变化的需求。采用量化和蒸馏等技术，以便减少对硬件的依赖、降低基础架构成本并减轻对环境的整体影响。

探索特定于行业的应用场景

生成式 AI 解决方案可用于多种用途。寻找包含大量特定于行业的 AI 用例库，且提供适用于推荐引擎、客户支持等应用场景的预制模板的解决方案，从而缩短上市时间。允许您使用特定于业务的数据对模型进行调优的工具能够提供更丰富的上下文信息，从而生成更准确且更相关的响应。

最大限度地发掘 AI 投资的价值

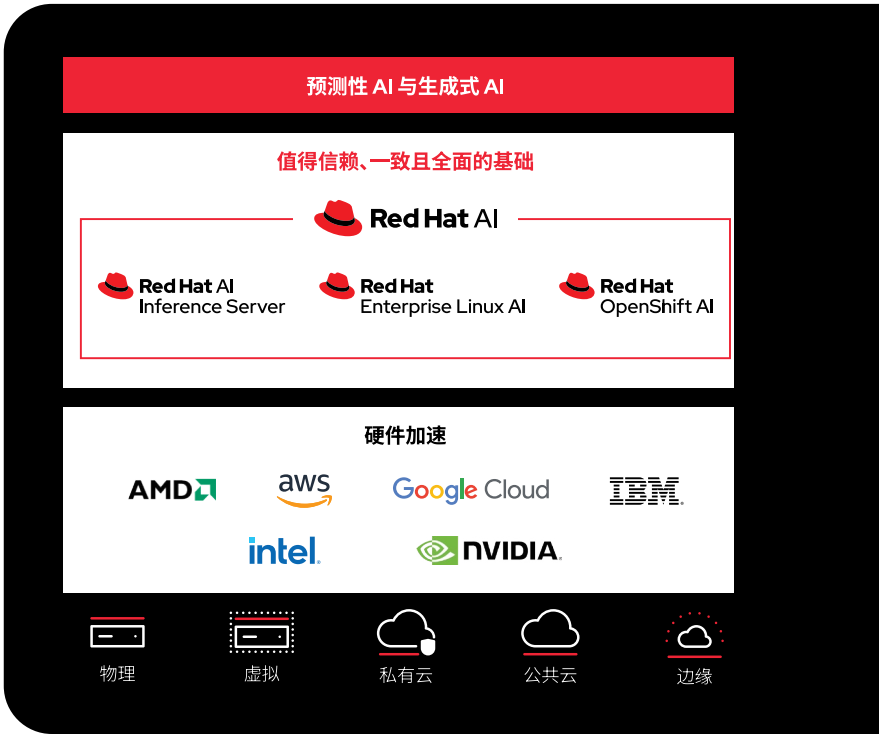
1. **使您的 AI 举措与业务目标保持一致。** 确保所选解决方案能直接支撑您的差异化、效率提升等战略目标。
2. **优化您的总拥有成本（TCO）。** 除了解决方案的前期成本外，还要考虑维护、基础架构和人才方面的开支。
3. **优先考虑采用程度和易用性。** 选择能在采用速度和团队实际能够用于提高工作效率的功能之间取得平衡的解决方案。
4. **利用集中式 AI 服务。** 通过设计并提供所有团队均可使用的可扩展模型即服务（MaaS），避免重复工作并优化 GPU 的利用。
5. **持续衡量和调整。** 通过衡量成本节省、效率提升和收入增长等因素来跟踪投资回报率（ROI），并相应地调整您的方法。

利用红帽 AI 缩短价值实现时间

红帽® AI 是一系列产品和服务组合，旨在加速混合云环境中 AI 解决方案的开发和部署。该组合专注于简化 AI 技术的采用，使先进的 AI 技术更易于在整个企业组织内推广应用。

通过在灵活性和一致性之间取得平衡，红帽 AI 可帮助您的团队在最适合您的工作负载和整体战略的环境中部署和管理**预测性** AI 模型及生成式 AI 模型。该组合可为您的 AI 采用之旅的各个阶段提供支持，从单服务器部署到高度分散的可扩展平台架构，因此，您可以从小规模开始，并根据自身需求和规划逐步扩展。支持各种硬件加速器、原始设备制造商（OEM）及云服务提供商，确保为您的 AI 工作负载提供稳定、经过优化且高性能的环境。此外，您还可以在包括本地基础架构、公共云和私有云资源及边缘位置在内的多种环境中部署 AI 应用和服务。

红帽 AI 产品组合包括：适用于单个 Linux 服务器环境的**红帽企业 Linux® AI**、适用于分布式 Kubernetes 平台的**红帽 OpenShift® AI**，以及用于优化大语言模型推理的**红帽 AI 推理服务器**。这些解决方案提供了开源技术和专用小语言模型（SLM），让您能够使用最新的 AI 工具，同时帮助您解决生成式 AI 通常涉及的成本高昂问题。事实上，随附的 Granite 系列模型（基于 Apache 2.0 许可证分发，训练数据集透明）能让您在更短的时间内开始使用生成式 AI，这些高效的模型可在保障性能的同时降低运维成本。与此同时，红帽企业 Linux AI 还提供产品技术支持和**模型知识产权（IP）保障**，助您有效规避风险，同时以透明可信、经济高效的方式专注于 AI 解决方案的构建、部署与管理工作。最后，**红帽 AI 合作伙伴生态系统**提供了一系列经过测试、受到支持且获得认证的产品和服务，助您加速创新，应对业务和技术层面的双重挑战。



简化 AI 模型交付流程

红帽 AI 使您的团队能够使用机密的企业数据构建预测性和生成式 AI 模型。该组合包含必要的工具、GPU 支持以及按需提供的自助式环境，能够提高敏捷性并减少对 IT 部门的依赖。通过使用一系列经过预优化的开源 Granite 系列模型，您可以高效地定制解决方案以满足特定的用例需求。这些平台通过集中管理模型、应用和代码来简化应用与 AI 模型的集成。红帽 AI 专为企业级生产工作流而设计，优先考虑安全性、成本优化和运维效率，通过治理、监控、安全保障、机器学习运维（MLOps）及大语言模型运维（LLMOps）服务提供可靠的日常支持。此外，支持在本地或私有云实例中进行气隙部署，从而降低敏感数据暴露的风险。

通过小语言模型（SLM）降低复杂性和成本

与大语言模型相比，小语言模型规模较小且所需的计算资源、数据和能耗也较少，是适用于许多应用的高效且具有成本效益的生成式 AI 模型。红帽 AI 产品中随附的专用开源 Granite 系列模型可帮助您控制 AI 成本并更轻松地上手使用。

我们还提供了一些工具，让您能够以注重安全的方式使用自己的企业数据对模型进行微调，从而在确保模型准确性和相关性的同时避免不必要的复杂性和成本。

阅读电子书可详细了解开源小语言模型的优势。

红帽 AI 的优势

提高效率

使用 Granite 系列模型及一系列经过预优化的开源模型有助于提高 AI 部署和运维效率。对这些模型进调优时所需的计算资源较少，同时可实现更快的推理速度，帮助您减少对硬件的依赖并最大限度地降低成本。

简便性和可访问性

面向各类角色（从开发人员到数据科学家再到 AI 工程师）的 AI 工具可加快模型的开发和定制流程。通过简化环境设置和优化模型训练与调优的硬件资源分配，红帽 AI 使企业级 AI 在整个企业组织中变得更加易于使用和部署。

灵活部署

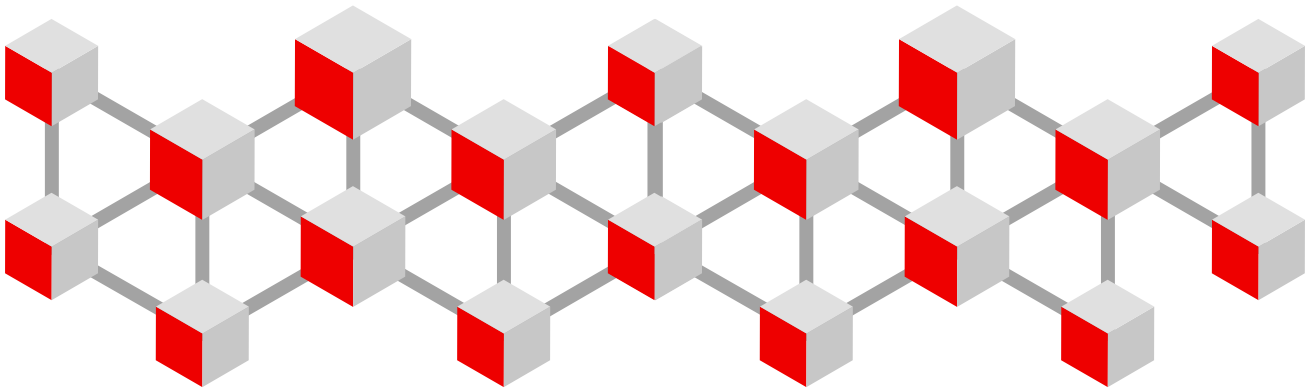
跨混合云环境的一致体验让您能够灵活选择模型及生成式 AI 应用的训练、调优、部署和运行位置。这有助于您遵守数据限制、保护隐私和维持安全性，同时控制 AI 基础架构的成本。

从实际应用和用例着手

您可以使用红帽 AI 产品组合来实施一系列 AI 用例以应对诸多业务挑战。一致的用户体验让从 AI 开发人员、数据科学家到 IT 运维团队的利益相关者能够更轻松地在混合云环境中开发和部署 AI 解决方案。

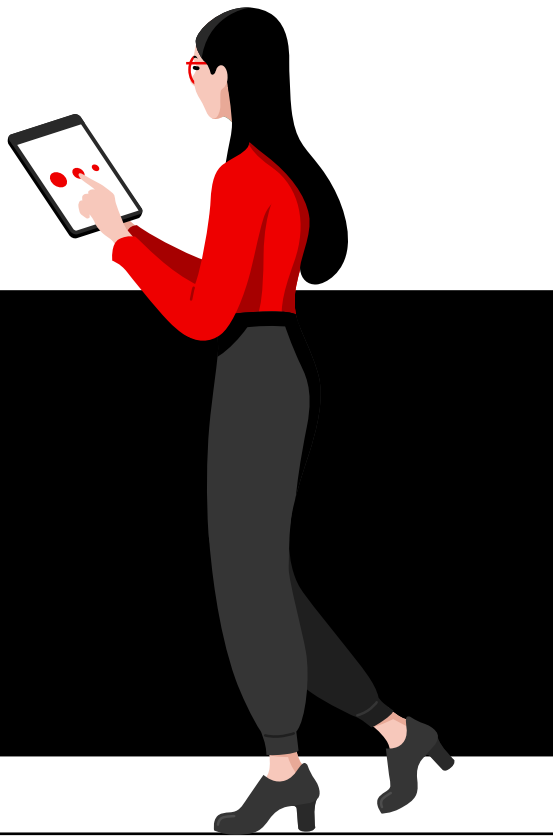
红帽 AI 可应对的常见业务用例

- 自然语言处理
- 内容创作
- 知识库
- 数字助手
- 媒体创作
- 服务个性化
- 推荐引擎
- 数据分析
- 网络安全
- 聊天机器人
- 任务和工作流自动化
- 情绪分析
- 计算机视觉
- 软件开发



基于 AI 和数据的业务运维

AI 模型能够处理企业组织收集的大量和广泛的数据，帮助他们做出更明智的业务决策。借助更深入的洞察，团队可以最大限度地提高收入、优化运维、改善客户体验并提高员工的工作效率。



了解红帽 AI 如何为您的企业组织助力

红帽提供了丰富的学习资料和工具，助您开启 AI 之旅。探索我们提供的面向业务领导者和技术学习者的 AI 学习路径。我们的分步课程涵盖了从 AI 基础知识到实操工具概览的内容。完成某个学习路径即可获得认证并提升您的 AI 技能。

开始学习

> 红帽 OpenShift AI 确保了[我们的研究人员和科学家] 拥有**搜索文本、搜索图像、训练模型**以及未来处理基因组数据所需的计算能力。
>
> **Eyal Dviri**
> Clalit Health Services 数据部门
> 创新团队主管

推荐引擎

AI 推荐引擎可基于历史数据对当前情况进行评估，以识别共同因素并提供指导建议。它们用于提供实时的行动建议，在许多行业均可使用。

Clalit Health Services 最近基于红帽 AI 建立了一个先进的 AI 平台，用于处理历史医疗数据并训练一个大语言模型，以识别可能需要预防保健和药物治疗的患者。然后，该解决方案通过与聊天机器人类似的交互方式为患者治疗提供行动建议。Clalit 还利用该平台构建学习流程和算法，以识别新趋势、患者行为模式和疾病规律等。

阅读成功案例

自助式的自动化 AI 工作流

AI 模型和应用的开发可能会很复杂。自动化 AI 管道和自助式操作可以简化这一过程，同时提高安全性和合规性。

DenizBank 的数据科学家希望将其现有工作流转变为一个手动操作更少、方法更标准化的流程。该银行的 IT 子公司 Intertech 提供了一个带有自动化管道和标准化规范的模型开发环境，以提高客户贷款识别与欺诈检测功能的开发效率，并加快推向市场的速度。其中一项关键改进是，Intertech 采用了红帽 AI，主要看重其自助服务功能以及在模型服务规模化和运维效率提升方面的优势。该银行的 100 多位数据科学家如今能够专注于构建比以往更稳健且更安全的模型。

阅读成功案例

> 凭借红帽 OpenShift AI 这款极具价值的 AI 驱动型解决方案，我们的数据科学家能够在一个精简的环境中**构建并部署更稳健且更安全的模型。**
>
> **Okan Çetinkaya**
> DenizBank 首席数据官兼首席分析官

自动化服务工单路由

公共部门和私营企业组织均广泛采用工单系统来为公众、客户和员工提供服务。基于 AI 的评估能够帮助这些组织将收到的工单快速路由至合适的团队。部分工单甚至可以自动处理，从而加快解决速度并提高用户满意度。

> AGESIC 正在使用 OpenShift 和 OpenShift AI，将**架构和软件开发方面的最佳实践**与治理流程相结合。
>
> **Gabriel Hernandez**
> AGESIC IT 与运维总监

乌拉圭电子政务与信息知识社会局（AGESIC）通过采用红帽 AI 来跨政府机构实现 AI 的拓展、规模化和标准化。该解决方案使 AGESIC 能够高效完成模型的构建、训练、调优和部署工作，并促进数据科学家、开发人员和 IT 运维团队之间的紧密协作。例如，AGESIC 构建并部署了一系列模型，用于每月将 2,000 条公民诉求自动分类并路由至合适的团队，将路由时间从一小时缩短至仅仅数秒。

阅读新闻稿

通过部署 4 个由 AI 驱动的支持解决方案，仅仅 10 个月内，红帽预计节省了约

150 万美元
的成本。

> AI 增强不仅能提高效率，**还能提升内容创作的质量**，且有助于提高工作满意度。
>
> **Mandy Elliott**
> 红帽 AI 与数据体验工程高级总监

客户支持和内容创作

优质的客户支持对于提供高价值的用户体验至关重要。AI 能够帮助支持团队在故障排查、生成信息摘要和工单摘要以及基于现有文档生成定制化内容等方面实现效率提升。

我们在**自己的企业组织**内部使用红帽 AI，以提升面向客户群体的客户支持和技术支持服务的效率和可扩展性。红帽的体验工程团队开发、测试并部署了 4 个由 AI 驱动的解决方案，这些解决方案的目标都是为客户和支持人员简化 IT 支持流程。这些工具可提高自助服务能力，提高效率，且有助于更快地响应支持案例。例如，我们提高了知识内容的可访问性，并最大限度地减少了每月要处理 30,000 个新案例的 IT 支持人员的重复性工作。我们由 AI 驱动的举措充分展现了这些解决方案的成本节省潜力：仅仅 10 个月内，我们就节省了约 150 万美元的支持成本，预计总体节省金额超过 500 万美元。

阅读成功案例

虚拟助手和聊天机器人

基于 AI 的聊天机器人和智能助手在响应质量和准确性方面持续提升。它们通常作为高级 AI 解决方案的交互入口，且广泛应用于从客户服务到信息传递再到内容创作的众多用例。

维也纳市希望提高市政员工的工作效率和满意度。该市开发了一款虚拟助手，旨在为员工的日常工作提供支持，通过即时解答与工作相关的问题，帮助他们更准确地回应市民的咨询和请求。借助红帽 OpenShift 上的 OpenShift AI，该市政府能够更快地开展创新，向公众提供新服务和新功能，并保持高频的版本迭代。

阅读成功案例

马上行动：与红帽携手，开启生成式 AI 之旅

构建适合您的 AI 解决方案。

利用您拥有的资源和独到的洞察，在充分自由的环境下，实现 AI 价值。红帽 AI 可加快推向市场的速度并降低在混合云环境中交付 AI 解决方案的运维成本。使用您自己的企业数据对小型、专用的模型高效调优，并获得在数据所在的任何位置进行部署的灵活性。规模化管理和监控 AI 模型的生命周期，并专注于利用 AI 进行创新以实现您的业务目标。

进一步了解红帽 AI 产品组合

了解产品信息、关键功能和优势，并免费试用红帽企业 Linux AI 和红帽 OpenShift AI 以及使用开发人员沙盒。

了解更多

探索面向您的整个企业组织的 AI 资源和知识

利用专为从业人员提供的实训资源来提升 AI 技能，或利用 AI 知识资源强化决策能力。我们提供演示、指南和案例研究，帮助您快速上手。

培养技能

版权所有 © 2025 Red Hat, Inc。红帽、红帽企业 Linux、红帽 Logo 和 OpenShift 是红帽在美国和其他国家/地区的商标或注册商标。
Linux® 是 Linus Torvalds 在美国和其他国家/地区的注册商标。所有其他商标均为其各自所有者的财产。

0425_KVM

Red Hat