

开启企业 AI 之旅： 新手指南



Red Hat

目录

简介

第 3 页

第 1 章

AI 的崛起

第 6 页

第 2 章

选择合适的 AI 模型

第 9 页

第 3 章

前期准备工作

第 12 页

第 4 章

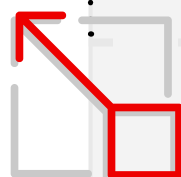
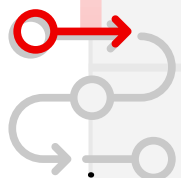
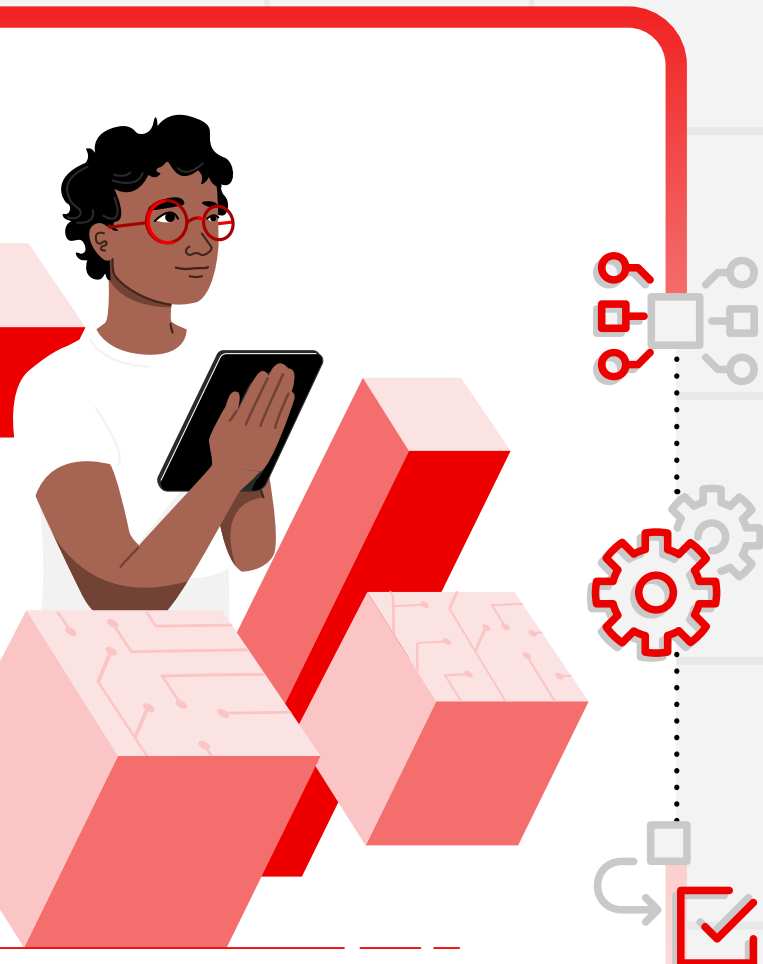
在红帽的帮助下 采用和扩展

第 16 页

了解更多

准备好在 AI 采用之旅 中迈出下一步了吗？

第 20 页



简介

企业组织逐渐认识到人工智能 (AI) 可为其企业的各个方面带来的机遇。

从客户接洽、支持和销售，到 IT 基础架构、流程、代码开发和解决方案交付，AI 的应用范围不断扩大，在各行各业蓬勃发展。

根据 IDC 的数据，预计到 2027 年，AI 市场规模将超过 4,230 亿美元，五年复合年增长率 (CAGR) 将达到 26.9%，并且许多企业将其 AI 计划的重点放在提升运维效率、改善客户体验和提高工作效率上¹。

在这种快速发展的形势下，领导者们的压力来自如何识别、选择、构建并交付能够为企业组织提供竞争优势的 AI 解决方案。但是，大多数企业组织的 AI 创新发展速度和提升 AI 成熟度的能力不匹配。这种情况导致难以释放 AI 的全部价值，而且在很多情况下，带来的问题数量多于能够解决的问题数量。



无论您是刚开启 AI 之旅，还是希望进一步了解 AI 对业务的影响，抑或是想知道如何扩大现有的 AI 实施规模，均可在本电子书中了解对与当今许多 AI 问题的解答。

¹ IDC FutureScape 网络研讨会。“[2024 年全球人工智能和自动化预测](#)”。文档编号：US51901124，2024 年 3 月。

AI 的类型有哪些？

要充分利用 AI，就必须全面了解它，其中包括当今企业组织正在使用的两种最主要的类型。

预测性 AI：预测性 AI 可利用历史数据，帮助企业组织识别模式，并做出符合未来需求的明智决策。预测模型可为需求预测、预测性维护和运维规划方面的应用提供支持。预测性 AI 以完善成熟的数据科学和机器学习（ML）技术为基础，使 AI 能够随着处理的数据量增加而不断改进。



生成式 AI (gen AI)：生成式 AI 由 Transformer 等深度学习模型提供支持，可以创建文本、图像和代码等新内容。它尤其适用于聊天机器人、自动化内容生成和创意工具等方面的应用。生成式预训练 Transformer (GPT) 等模型通过生成类人文本和图像，在自然语言处理和创意领域掀起了一场革命。

实施 AI 的益处有哪些？

AI 的全部潜能仍有待发掘，但了解这项快速发展的技术如何使各行各业中不同规模的企业组织受益，是确定从何处着手将 AI 整合到您企业组织中的好方法。

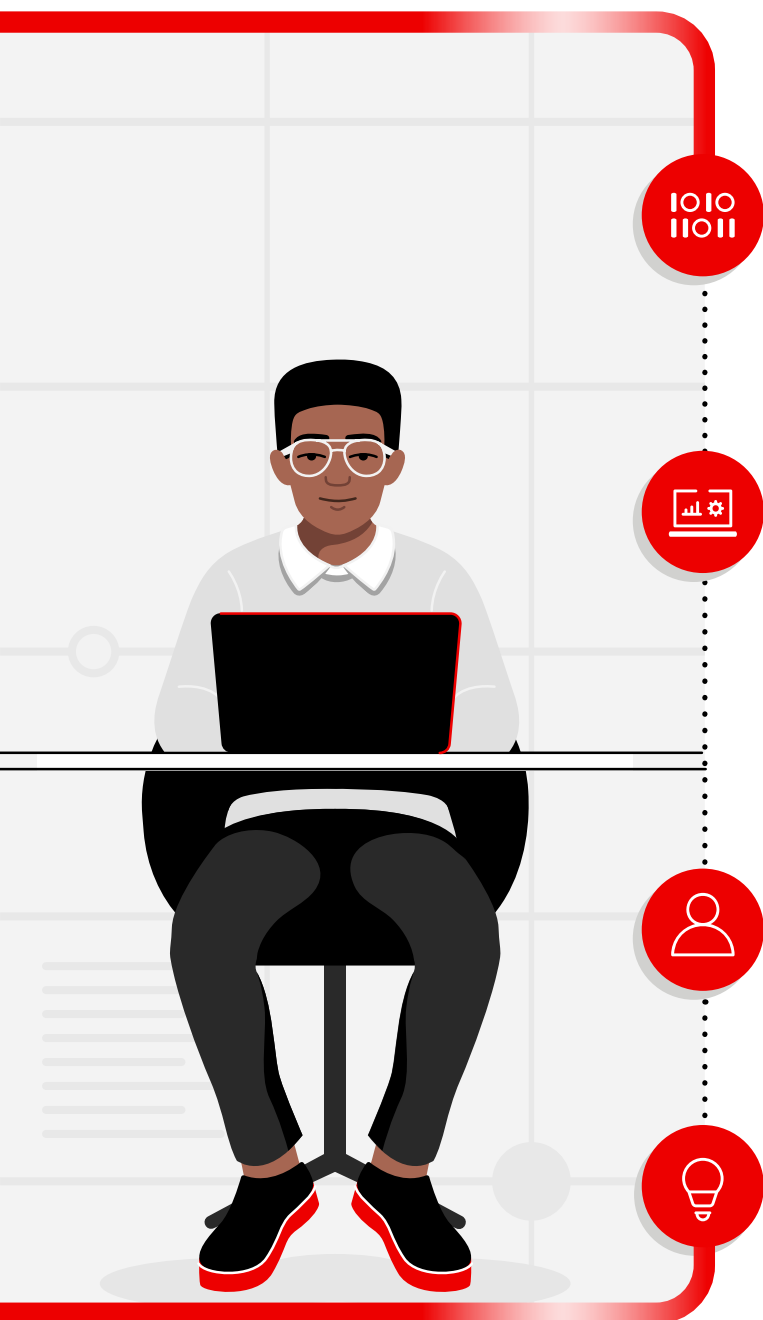
考虑 AI 的以下优势，以及这些优势如何使您的企业组织受益：

数据量。随着数据的指数级增长，企业组织往往难以管理他们收集到的大量信息并从中获得深刻的见解。AI 可以快速处理和分析大型数据集，发现人们难以识别的宝贵见解和趋势。

运维效率欠佳。许多企业组织都明白，低效的流程和瓶颈会阻碍工作效率的提高，导致需要花费更多的时间和精力来消除这些障碍。将自动化与 AI 结合使用有助于简化运维，从而减少错误并提高流程效率。例如，这包括一些简单的应用，比如自动生成包含行动项目和明确的后续步骤的会议记录，或加速网站或社交媒体的图形和视频创建。

客户的预期。客户希望获得无忧的个性化体验。通过分析客户数据并提供量身定制的建议和互动体验，AI 可以改善客户服务并提高个性化水平。

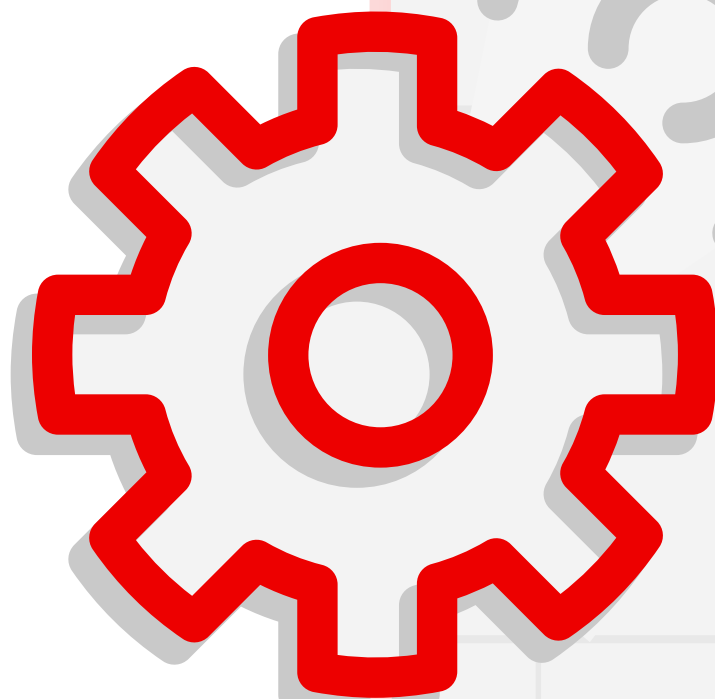
市场竞争力。要在瞬息万变的市场上保持竞争力，就需要不断的创新。AI 可以帮助企业组织快速适应市场变化，保持竞争优势。生成式 AI 甚至可以在领导层使用它来寻找灵感时或者在准备重要会议时用于完善具体方法。



AI 的崛起

AI 已经发展了数十年，推动了医疗卫生、金融和制造等行业的进步。

不过，最近大热的生成式 AI 因能够生成类人文本、逼真图像甚至软件代码而备受关注。与自动执行任务或分析数据的传统 AI 不同，生成式 AI 能够创造性地解决问题和生成高级内容。



有助于加速创新的 AI 模型类型

大语言模型（LLM）和稳定扩散模型是促使生成式 AI 出现爆炸性增长的 AI 模型之一。GPT 等 LLM 基于海量数据集进行预训练，能够理解和生成自然语言，因此在客户支持自动化、营销文案生成等方面非常有价值。另一方面，稳定扩散模型可以创建超逼真的图像，推动娱乐、营销等领域的创新。

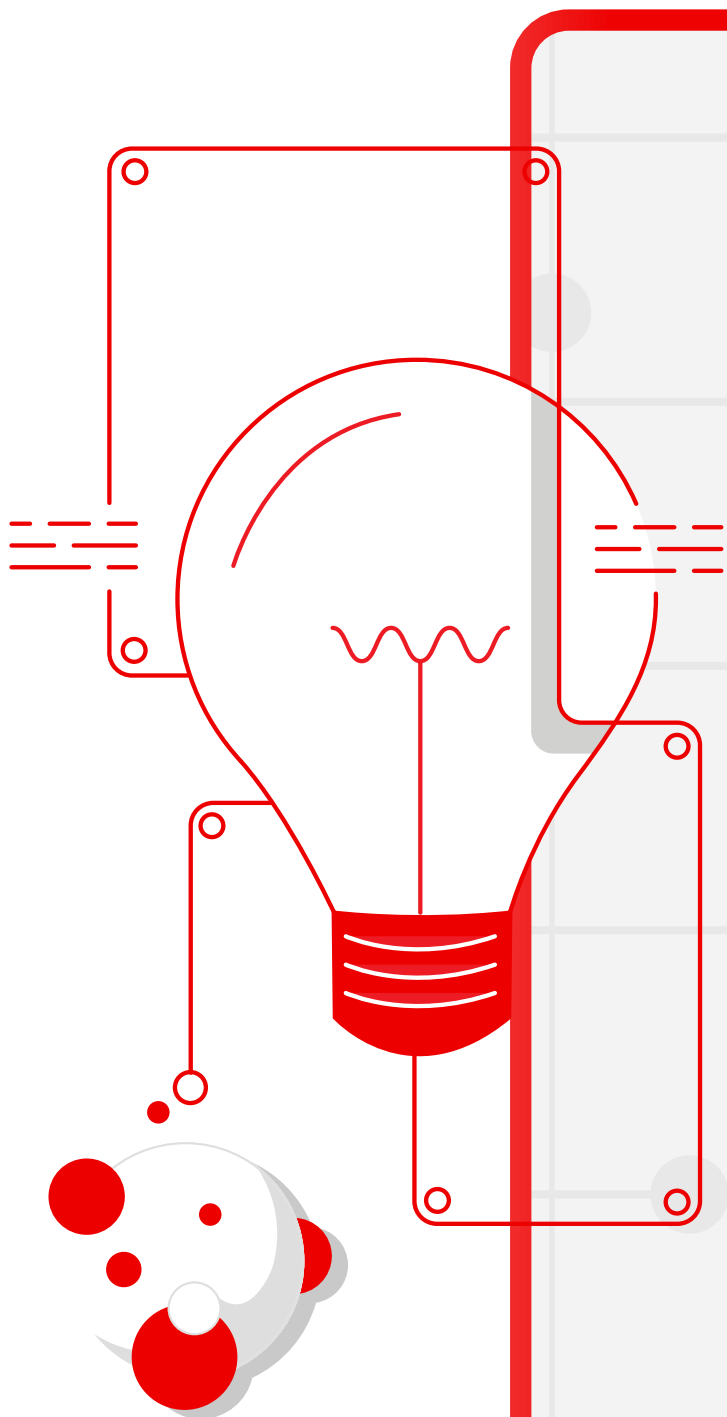
不容忽视的新兴趋势

企业组织越来越多地探索多模态 AI，此类 AI 将文本、图像和数据处理功能整合到一个模型中，提供更全面的解决方案。要在企业内充分发挥 AI 的潜力，关键是要及时把握这些趋势。

开源：AI 创新的基础

红帽的 AI 战略深深植根于开源，可帮助企业在透明、可信和成本更低的情况下推进生成式 AI 发展。通过使用红帽的开放式**混合云**平台，企业组织可以自由地进行创新，同时保持对其 AI 解决方案的掌控力。

[进一步了解 LLM 及其运作方式](#)



通过开源方法控制 LLM

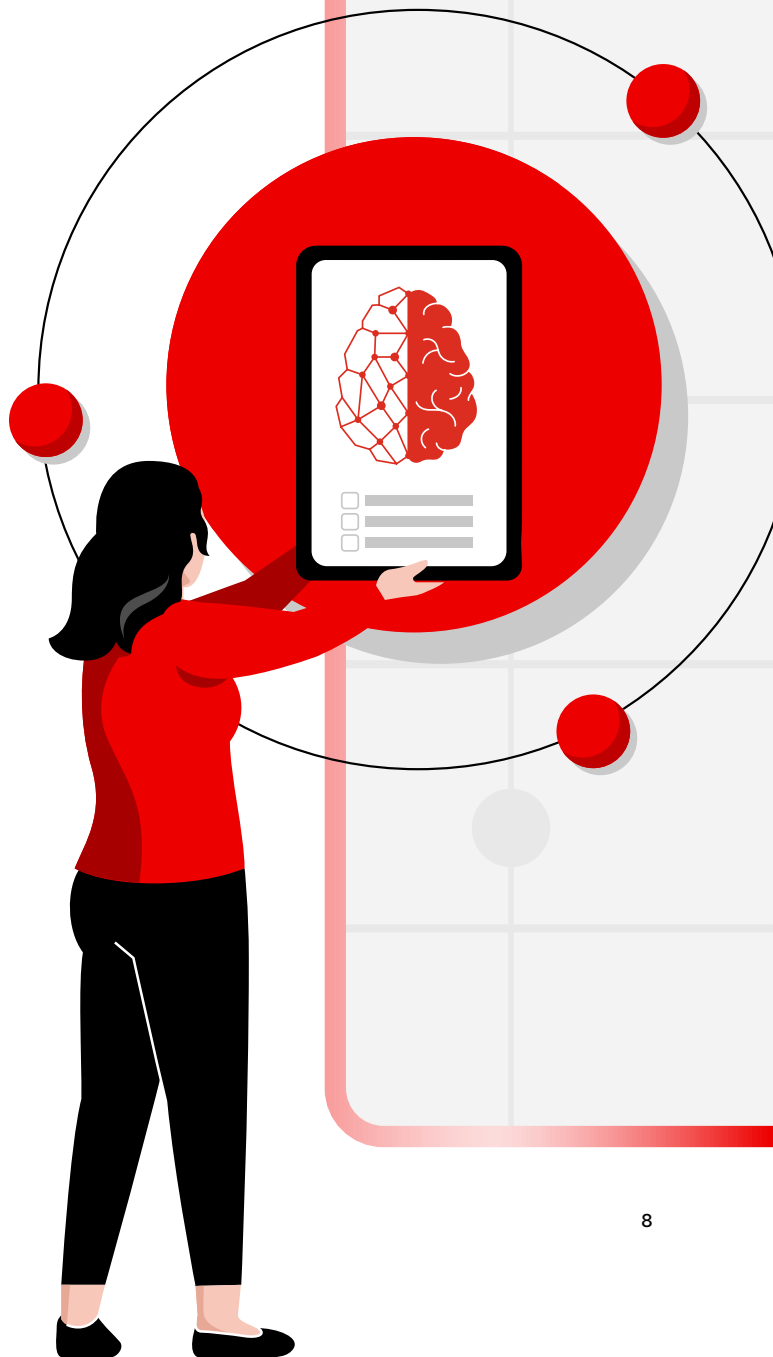
虽然生成式 AI 正在改变商业领域的几乎各个方面（从软件的开发方式到我们的沟通方式），但用于实现生成式 AI 强大功能的模型（LLM 及其他类型）通常受到服务提供商的严格控制。这就意味着，企业需要具备专业技能，有时还要付出高昂的成本（金钱和时间），否则难以评估生成式 AI 服务的能力。

如果企业没有详细了解创建模型的数据集或模型使用数据的方式，那么企业将面临 AI 生成的内容方面的潜在风险。如果根据拥有版权的源代码训练代码生成模型呢？使用该模型生成的所有代码是否也属于受版权保护的代码？很多类似问题还没有明确的答案，但在了解到后果可能会很严重后，企业纷纷转为采用开源 AI。

红帽的 AI 方法植根于开源，我们对 [IBM Granite 系列](#) 基础模型等开源模型的支持也是如此。

红帽的 AI 解决方案甚至可以通过 [InstructLab](#)（一种可增强 LLM 能力的社区驱动型解决方案）直接为 AI 模型的开发做贡献。

[在 GitHub 上探索 InstructLab](#)



选择合适的 AI 模型

不同的 AI 模型可用于多种不同的用例。

预测性 AI 和生成式 AI 模型均可在单一应用或服务中使用。每种模型的费用和优势各不相同，但都可以帮助企业组织缩短上市时间以完成初步概念验证。图像分割、语音转文本和图像识别模型都是功能强大的常见模型，但关键在于评估哪种模型最适合您的用例。

基础模型是基于大量数据训练出来的，功能灵活多变，但由于规模庞大，可能会导致成本增加、管理要求提高并且复杂性增加，这意味着此类模型可能并不适合所有用例。

如果您希望对某个现有模型进行调优以满足您的要求，那么经过微调的小型模型（仍属于生成式 AI 系列）可能是更理想的解决方案。在决定选择何种模型时，您可以选择预构建的模型，这种模型易于获取和使用，也能很轻松集成到您的系统中。此类模型的一个常见示例是 LLM，这是一种强大的工具，已使用大量数据进行训练。

但是，如果您有特定的业务要求、数据隐私方面的顾虑，或者希望对模型的行为有更大的控制权，则可能需要构建并自行托管一个定制模型。

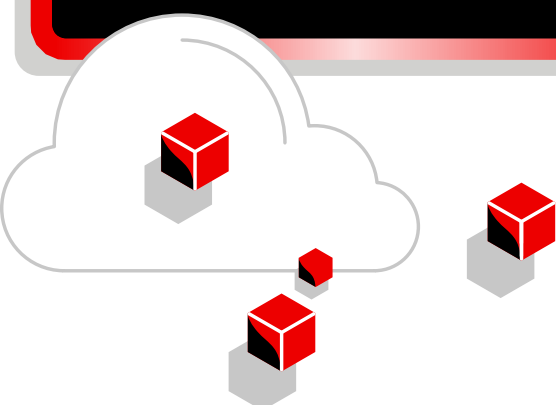
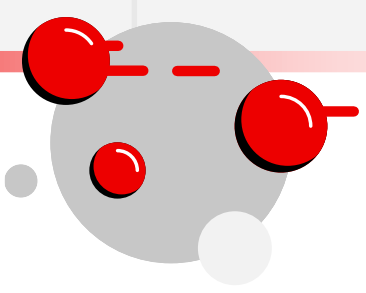
模型构建与模型调优

从零开始构建 AI 模型是一项艰巨的任务。您需要收集和准备与企业组织的业务用例相关的大型数据集。然后，您必须选择合适的算法，并使用这些数据对其进行训练。此过程需要强大的计算能力和丰富的专业知识，因此是一项耗时耗力的工作。虽然构建传统或基础模型可为用户提供自定义解决方案，但这并非始终是最有效的途径。

另一方面，要对基础模型进行调优，需要根据具体要求调整预训练的模型。一种常见的方法是迁移学习，即使用大型数据集训练过的模型，在特定领域的较小数据集上对其进行再训练。这种方法可以让模型保留在初始训练中学到的一般知识，同时根据特定数据的细微差别进行调整。

微调模型

另一种方法是微调模型，即调整模型的参数，以提升针对特定任务的性能。模型参数是指选定模型中的变量，可以通过将给定数据拟合到模型中来进行估算。微调可能涉及更改学习率、修改模型的架构或对模型中的某些层进行更加密集的训练。这些方法有助于强化模型的知识，使其对特定用例更加有效。



InstructLab 采用微调方法，目的是降低对必备 AI 知识的要求，简化针对现有 Granite 基础模型的企业知识补充。

微调模型的替代方案

研究人员正在研究如何调优基础模型，以提高速度和效率。检索增强生成 (RAG) 就是一种常见的方法，可从已对额外知识（背景信息）进行了编码的外部来源检索事实。

RAG 主要使用一个或多个外部数据库（向量数据库），为生成式 AI 提出的问题提供额外的背景信息。另一种新兴方法是代理式 AI 系统，它可将多个生成式 AI 代理整合到一起，从外部系统（例如内部数据库、企业内联网或互联网）查询知识，为生成式 AI 模型提供最准确的最新信息。

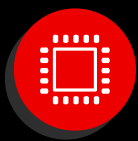
最后一个示例是提示词调优，即 AI 模型接收提示或前端提示词（包括额外的词语或 AI 生成的数字），引导模型做出所需的决策。RAG 查询的结果将为提示词提供额外的背景信息，并且提示词调优和 RAG 协同发挥作用，提供更加出色的响应。通过对现有 LLM 进行微调、使用 RAG 方法以及对提示词进行调优，数据有限的企业组织可以针对特定任务量身定制基础模型。

支持 AI 模型的基础架构与模型本身同样重要。不同的任务需要不同类型的硬件。



中央处理单元 (CPU)

处理一般计算任务的传统处理器。它们用途广泛，但对于大规模 AI 工作负载来说可能并不高效。



图形处理单元 (GPU)

专为处理并行处理任务而设计的专用处理器，是训练需要同时处理大量数据的深度学习模型的理想选择。



神经处理单元 (NPU)

一种专为 AI 任务设计的新型处理器，可为某些类型的模型提供更高的效率和速度。

混合云在企业采用 AI 方面的作用

混合云环境在企业采用 AI 的过程中发挥着至关重要的作用。混合云将本地基础架构与公共云和私有云资源相结合，让您能够灵活选择如何以及何时部署和管理 AI 工作负载。例如，您可以使用功能强大的云端 GPU 来训练 AI 模型，然后出于安全性或合规性方面的考虑，将模型部署到本地或私有云中。因此，在利用混合云方法时，一个关键的考量因素是所选工具和平台的一致性。

红帽的开放式混合云方法有助于企业组织跨不同环境集成 AI，从而提高一致性、可扩展性和灵活性。借助此方法，您可以跨多个云环境、本地或网络边缘管理 AI 工作负载，优化数据放置并促进数据平稳迁移，从而更轻松地在企业范围内采用 AI。

通过了解 AI 模型、数据和基础架构，您可以更好地应对 AI 采用的复杂性并充分发挥它的潜力。

前期准备工作

与采用任何新技术一样，采用 AI 的过程中也面临着一些挑战，要想取得成功，企业组织必须克服这些挑战。

利用以下考量因素来评估您的企业组织的就绪情况，并确定您可能需要重点关注的领域，以加快 AI 的采用。

评估数据质量和可用性。对于 AI 来说，能够访问相关的优质数据至关重要。数据质量对于训练准确的 AI 模型至关重要，因此很有必要评估数据的完整性、准确性和相关性。

评估技术基础架构。确定您当前的基础架构能否支持 AI 工作负载。其中包括评估高性能计算资源、存储解决方案、自动化和网络功能的可用性。

确定哪些方面需要技能。评估您的企业组织内 AI 专业知识的可用性。评估当前的技能组合，确定哪些方面需要培训技能或专业技能。

审查战略一致性。确保您的 AI 计划与您的业务目标和战略保持一致。AI 项目应支持企业组织的总体战略目标，并提供可衡量的业务价值。



如何开启 AI 之旅

企业组织内 AI 采用的速度和规模取决于多种因素，但对于几乎所有技术现代化项目来说，先在小范围内试用并逐步扩大应用范围往往是一种不错的方法。

以下 8 个步骤可帮助您企业组织开始并推进 AI 采用之旅：

1

评估能力和目标

首先要评估您的企业组织当前的能力、基础架构和战略目标。确定 AI 是否与您的企业组织更广泛的目标一致，并识别 AI 可带来更多价值的潜在领域。这一初步评估将有助于为您的 AI 采用之旅设定明确的方向。

2

确定用例和 AI 团队

确定您的企业组织内 AI 可以处理的业务机会。组建一个专门的 AI 应用团队，吸纳跨职能成员（例如开发人员、领域专家、数据科学家和 IT 专家）来领导这项计划。明确定义的用例将会指导您的 AI 采用工作并将资源集中起来。

3

模型选择

根据已确定的用例选择合适的 AI 模型。无论是用于生成式 AI 的 LLM，还是用于数据分析的预测模型，都要确保模型的功能与用例的目标保持一致。考虑模型的复杂性、可扩展性以及现有系统的兼容性等因素。

4

测试和验证循环

明确 AI 实施的成功标准，例如性能指标、准确率或业务目标。建立测试和验证循环，持续评估模型的有效性。根据这些循环定期提供的反馈，您可以对模型进行微调，并确保您的 AI 之旅顺利推进。

5

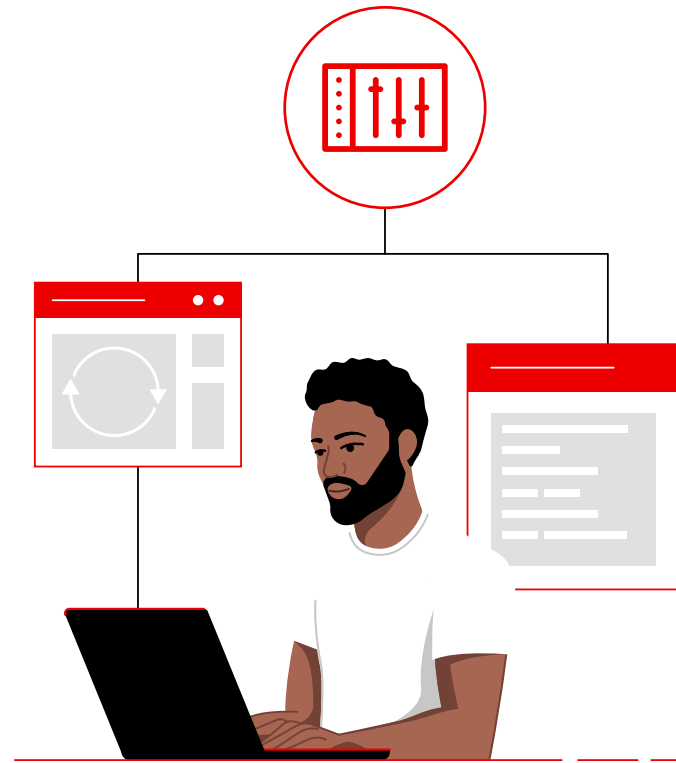
模型调优

使用您的企业组织的数据来自定义所选模型。此调优流程涉及向模型提供相关数据，以提高模型的准确性和与特定业务用例的相关性。通过对模型进行微调，可确保模型适应您的企业组织的独特环境且能满足需求。

6

合成数据训练

建议使用合成数据来进一步训练和强化 AI 模型。这种方法利用 LLM 教师和 LLM 学生等方法，可让您在真实数据稀缺或敏感的情况下生成优质训练数据。合成数据有助于提高模型的鲁棒性和性能，同时又能确保隐私安全。



7

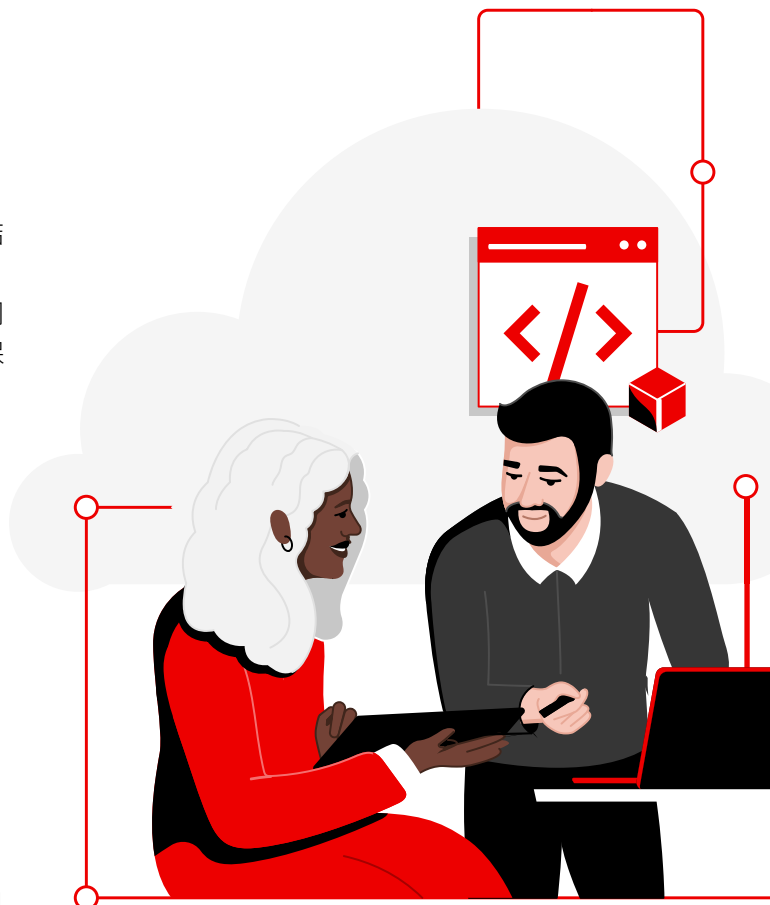
监控偏差

偏差监控提供基于内容的一般监控。偏差监控不是结构化的配置管理，而是跟踪本地文件系统内容的变化。偏差监控可帮助您监测和处理模型准确性随时间推移而出现的任何偏差或下降问题。持续监控可确保模型在条件发生变化时保持有效性和相关性。

8

向专家寻求帮助

如果您的内部 AI 专业知识仍在不断积累的过程中，那么向外部专家（例如红帽® 咨询）寻求帮助将是一项巨大的优势。红帽专家可指导您解决 AI 采用方面的复杂问题，提供有价值的见解，并为您提供培训机会。红帽咨询可加快您的 AI 之旅，提高成功几率。



AI 的采用取决于协作

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²。

- **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。
- **开发团队和社区**需要从一开始就参与进来，合作、创建、共享和改进开源工具、框架和最佳实践，以促进 AI 的采用。这将确保 AI 的使用与业务价值相关联。

² Kearney. “[组建强大的团队以应对生成式 AI 的复杂性](#)”，2023 年 11 月 15 日。

第 4 章

在红帽的帮助下 采用和扩展

红帽 AI 能够在混合云环境下确保用户信任、提供多种选择并保持一致性，以加快企业对 AI 的采用。

红帽 AI 产品组合包含红帽企业 Linux® AI（适用于单个 Linux 服务器环境）和红帽 OpenShift® AI（适用于分布式 Kubernetes 平台环境），且集成了机器学习运维功能。这两种解决方案均基于开源技术和模型，可帮助企业组织始终走在 AI 创新的前沿，加快探索步伐，让更多人可以轻松使用先进的工具和技术。

红帽广泛的合作伙伴生态系统可进一步增强您的 AI 能力。例如，以普及 GPU 而闻名的主要 AI 公司英伟达继续与红帽合作，通过提供针对 AI 工作负载优化的端到端企业平台，充分发挥 AI 的作用。英伟达帮助企业客户在 AI 和高性能计算应用中采用 GPU 加速计算。

“红帽与英伟达有着长期的紧密合作关系，红帽企业 Linux AI 展现了我们共同的关注点，即为构建下一波 AI 技术和应用的开发人员和研究人员提供全栈计算和软件³。”

Justin Boitano，英伟达企业产品副总裁

³ 红帽新闻稿。 [“红帽企业 Linux AI 赋能无障碍的开源生成式 AI 创新”](#) 2024 年 5 月 7 日。



进一步了解红帽企业 Linux AI

红帽企业 Linux AI 包含 4 个基础组件：

1

开源 Granite 模型

红帽企业 Linux AI 包含红帽完全支持的开源 Granite 模型。借助这些灵活的模型，您可以创建自定义语言模型，并在公开或私有环境下使用它们。

2

InstructLab 模型一致性

InstructLab 是由红帽和 IBM 主导的开源项目。它利用特定知识自定义 AI 模型，并生成用于训练的合成数据。作为一款与 Git 存储库集成的命令行工具，用户可以轻松增加技能和训练模型。

3

针对 Granite 模型和 InstructLab 优化的可引导红帽企业 Linux

Granite 模型和 InstructLab 工具在专门针对 AI 优化的红帽企业 Linux 镜像上运行，该镜像几乎与所有硬件和云环境兼容。这种设置允许通过高端 GPU 实现高效性能，这是快速训练和模型部署所必需的。

4

企业支持和保障

红帽企业 Linux AI 订阅包含企业支持、从 Granite 7B 模型和软件开始的完整产品生命周期以及红帽的知识产权保障。

红帽企业 Linux AI 有助于将生成式 AI 应用投入使用

对于刚刚开始使用生成式 AI 的企业组织，红帽企业 Linux AI 在单一服务器开发和推理环境中提供了即用型 LLM 和代码语言模型。

这提供了一个包含模型和工具的统一环境，使您可以轻松开始使用生成式 AI，并使用业务数据自定义模型，而无需大量的 AI 专业知识或基础架构。

红帽企业 Linux AI 由红帽提供全面支持和保障，可降低风险。它还提供了一种简化的生成式 AI 实施方法，更方便开发人员和领域专家使用，并且可让他们互相协作，更快看到业务成效。

为何选择红帽企业 Linux AI?

面向企业的 LLM

具有开源许可的 IBM Granite LLM 使用 Apache-2.0 许可证，可获享红帽的全面支持和保障。

社区协作

InstructLab 可简化生成式 AI 模型的实验和对齐调整。

云原生可扩展性

红帽企业 Linux 镜像模式可将 AI 平台作为容器镜像进行管理，从而简化您的扩展方式。

加速和 AI 工具

开源硬件加速器以及经过优化的深度学习功能有助于更快地取得成效。

利用红帽 OpenShift AI 进行扩展

红帽 OpenShift AI 是一个集成式 MLOps 平台，用于在混合云环境中大规模构建、训练、调优、部署并监控依托 AI 技术的应用、预测模型和基础模型。

红帽 OpenShift AI 在红帽 OpenShift 的基础之上构建而成，可在处理 AI/ML 项目的工作负载和性能需求时提供一致且简化的自动化体验。MLOps 实践有助于企业组织快速响应 AI 创新，更快地将依托 AI 技术的应用交付到生产环境中。

在红帽 OpenShift AI 沙盒中进行试验

借助 Open Data Hub 及其他开源项目中的精选组件，红帽 OpenShift AI 为数据科学家和开发人员提供了一个强大的开放式混合 AI/ML 平台，能够高效收集数据洞见并构建依托 AI 技术的应用。

[在开发人员沙盒中试用 >](#)

为何选择红帽 OpenShift AI?

扩展模型供应

您可以将供应的模型集成到在本地、公共云中或边缘运行的智能应用中，还可以基于对源 notebook 的更改来重新构建、重新部署和监控这些模型。

减少在 AI 基础架构管理上花费的时间

为您的团队提供对资源的按需访问权限，以便他们专注于探索数据并构建能够为您的企业组织带来更多价值的应用。

经过测试、享受支持的 AI/ML 工具

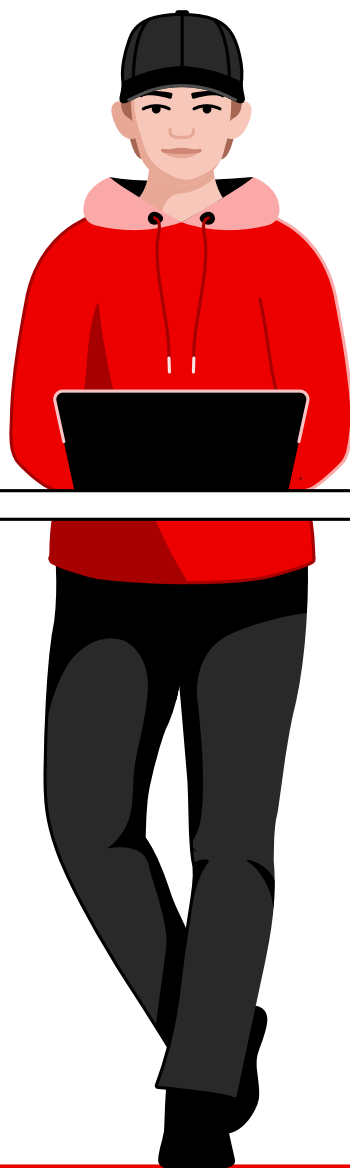
红帽平台可跟踪、集成、测试和支持在红帽 OpenShift 应用平台上运行的常见 AI/ML 工具和模型。

跨混合云的灵活性

红帽 OpenShift AI 既可以作为自助管理软件使用，也可作为红帽 OpenShift 上的全托管云服务提供，借助这个既注重安全又灵活的平台，您可以选择在任何位置开发和部署模型，无论是在本地、公共云中还是在边缘。

采用最佳实践进行运维

红帽咨询提供各种服务来协助您安装、配置和使用红帽 OpenShift AI，发挥这款产品的最大潜力。无论您想试用红帽 OpenShift AI Pilot，还是需要 MLOps 基础构建方面的帮助，红帽咨询都能提供支持和指导。



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。您能够获得坚实的基础，以开发和部署生成式 AI 模型，并针对快速采用提供服务的培训。

了解更多

准备好在 AI 采用之旅中迈出下一步了吗？

利用红帽的开放式混合云战略加快 AI 采用速度，灵活地在您需要的任何地方运行 AI 应用。

利用红帽的专业知识、咨询和培训服务，助力 AI/ML 项目启动，帮助您的企业组织实现 AI 愿景。

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

进一步了解红帽 AI



[进一步了解红帽企业 Linux AI](#)



[探索如何利用红帽 OpenShift AI 进行扩展](#)