

# 构建生产就绪型 AI 环境的首要 考虑因素



Red Hat

# 目录



数据是一项重要的  
商业资产

第 3 页



容器和容器编排

第 9 页



混合云平台

第 13 页



大规模 AI 推理

第 15 页



构建开放、灵活的  
AI 基础

第 18 页



合作伙伴亮点

第 21 页



准备好从您的数据中  
获得更大价值了吗？

第 23 页



构建生产就绪型 AI 环境

第 6 页



应用管理和 GenAIOps

第 11 页



模型自定义和对齐

第 14 页



AI 安全防护

第 17 页



利用经过认证的 AI/ML  
合作伙伴生态系统获得  
更多选择和更大的灵活性

第 20 页



成功案例

第 22 页

# 数据是一项重要的商业资产

1010  
11011

## 企业 AI 市场现状

生成式人工智能（生成式 AI）已跨越实验阶段，成为许多企业组织的日常工具。

团队利用它来总结内容、辅助编写代码和创作内容，并以更自然的方式与数据交互。在企业层面，领导者希望生成式 AI 能够帮助他们改善客户、员工和整体运营的成效，而不仅仅是回答零散的问题或制作有趣的表情包。

基于企业组织现有的数据和应用，生成式 AI 可以帮助企业组织实现以下目标：



将大量的非结构化内容转化为可搜索且可重复使用的知识。



协助开发人员、分析师和内容作者更快地创建和优化代码、报告及内容。



跨渠道为客户和员工提供个性化的数字体验。

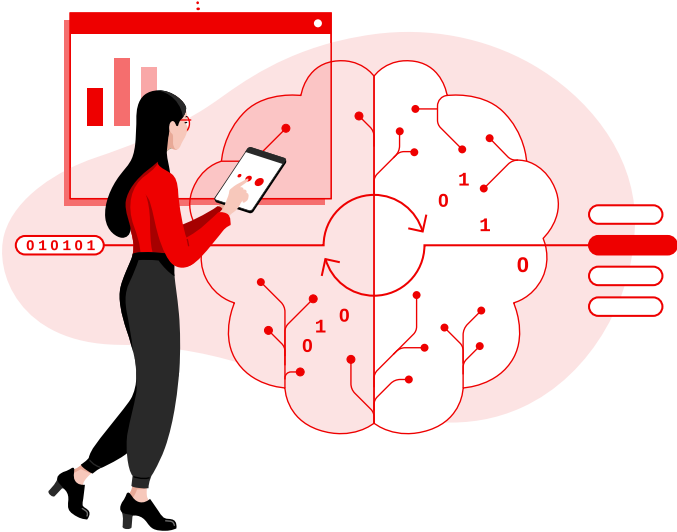


自动化处理遵循明确策略的常规决策和工作流。



提升开发、运营和业务团队的生产力。

近期行业研究表明，这种转变已经开始。IDC 报告称，超过半数的受访企业组织已在生产环境中运行多种生成式 AI 增强型应用或服务，并预计 2025 年至 2029 年间，AI 领域的支出将同比增长约三分之一，到 2029 年相关支出将达到约 1.3 万亿美元。<sup>1</sup>对于大多数企业而言，生成式 AI 正逐渐成为其核心产品与服务的一部分。



<sup>1</sup> IDC 白皮书。 “[Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and \\$1.3 Trillion in 2029, According to IDC](#)”（根据 IDC 的数据，代理式 AI 将在未来五年内主导 IT 预算增长，到 2029 年，代理式 AI 相关支出将占全球 IT 支出的 26% 以上，达到 1.3 万亿美元），2025 年 8 月 26 日。



与此同时，许多企业组织已将目光投向下一步：代理式 AI。代理式 AI 并未将生成式 AI 视为单一的聊天机器人或助手，而是使用可以调用工具、与应用交互以及协调多步骤任务的 AI 代理。在实践中，这种方法可以改变您构建和运维软件的方式，从客户自助服务和 IT 运维到复杂的业务 workflow，皆涵盖其中。

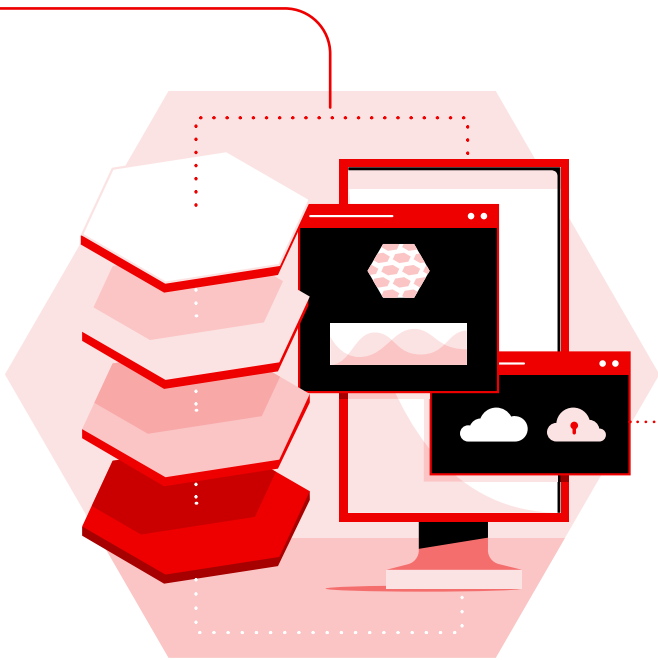
IDC 报告称，超过半数的企业组织已经开始真刀真枪地测试和试用代理式 AI，到 2026 年底，近三分之一的 AI 赋能应用将依赖于此项技术。<sup>2</sup>企业目前将其视为一条战略性的前进路径。

要想把握这一价值，您需要能够灵活地选择 AI 的运行方式和位置。

许多企业组织目前正规划混合 AI 基础架构，将公共云与专用的本地环境相结合。IDC 指出，公共云与本地基础架构的混合模式已成为最常见的数字基础架构策略，并且大多数决策者认为他们的 AI 工作负载需要采用混合部署方式。<sup>3</sup>

借助混合开放平台，企业组织可实现以下目标：

-  确保敏感数据与模型处于掌控之中。
-  满足数据隐私和主权方面的要求。
-  从一系列硬件选项中自由选择。
-  从广泛的开源模型中自由选择。
-  仍可在需要时利用云级扩展能力。



本电子书将介绍构建生产就绪型 AI 平台的核心步骤、企业组织在此过程中将会遇到的关键注意事项，以及红帽® AI Enterprise 如何提供统一的解决方案来助力构建该平台。

<sup>2</sup> IDC 白皮书。“Agentic AI Impact on Digital Infrastructure Strategies”（代理式 AI 对数字化基础架构策略的影响），文档编号 US53418526，2025 年 10 月。（需要购买）

<sup>3</sup> IDC 白皮书。“AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025”（AI 需求推动本地基础架构部署及与公共云互操作性的需求增长（2025）），文档编号 US53418426，2025 年 10 月。（需要购买）

# 各个行业的 AI 用例



## 医疗卫生

- 提升临床诊疗效率。
- 提高诊断速度和准确性。
- 改善患者治疗效果。



## 电信

- 深入了解客户行为。
- 改善客户体验。
- 优化 5G 网络性能。



## 保险

- 自动化理赔处理。
- 提供按实际使用情况计费的保险服务。
- 协助进行风险计算。



## 金融服务

- 个性化客户服务。
- 改进风险分析。
- 检测欺诈与洗钱行为。



## 汽车

- 支持自动驾驶。
- 预测维护需求。
- 优化供应链。



## 能源

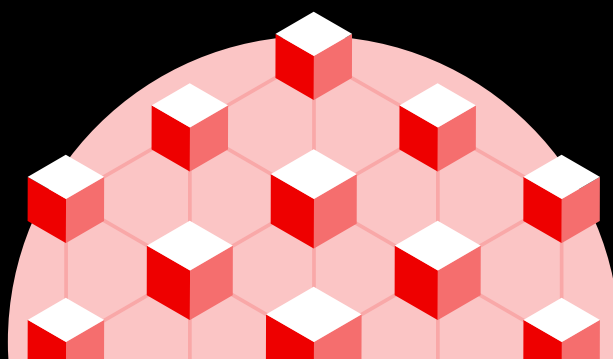
- 预测维护需求。
- 优化现场作业和安全保障。
- 加速油藏模拟和预测。

# 企业 AI 的构建模块

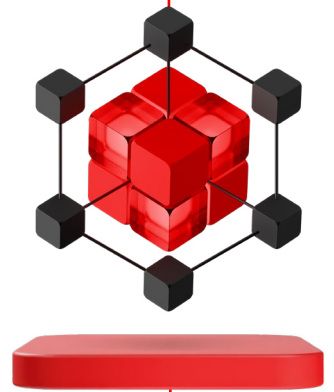
本电子书将探讨不同类型的 AI 如何在企业架构中协同工作。

- **生成式 AI:** 利用大语言模型 (LLM) 根据数据和提示生成文本、代码和其他内容, 使团队可以更快地工作并更轻松地进行实验。
- **预测性 AI:** 利用历史和实时数据来预估未来结果, 如需求、风险或设备健康状况, 使企业组织能够更早、更有信心地采取行动。
- **代理式 AI:** 使用 AI 代理, 这些代理可以调用工具、连接到应用, 并协调多步骤工作流以实现目标, 而不仅仅是回答单个问题。

- **AI 推理:** AI 推理是生产运行时阶段, 在此阶段, 模型将所学知识应用到新的真实数据中, 以返回预测、建议或操作。推理可以在混合环境中运行: 本地、云端或边缘。



# 构建生产就绪型 AI 环境



构建生成式 AI 驱动的应用和 AI 代理是一个重复循环、逐步改进的过程，而不仅仅是简单地创建 AI 模型。AI/ML 生命周期中的主要步骤如下：

- 1 确定您的用例，为您的 AI 计划设定业务目标，并获得利益相关者和领导层的支持。
- 2 选择模型实验和部署平台的运行位置：本地或云端。
- 3 选择最符合您需求的 AI 模型。选择开放模型，以避免锁定。
- 4 使用检索增强生成（RAG），根据您的专有数据自定义或调整所选模型。
- 5 在推理服务器中部署您的模型。
- 6 构建生成式 AI 驱动的应用或工作负载。
- 7 构建好工作环境后，即可通过代理式 AI 扩展和自动化 workflows。
- 8 以安全至上的方式大规模监控和管理模型。



借助开放且适应性强的 AI 架构，您能够更有效地执行这一流程。这种架构需要以下几种关键技术和能力：

- **访问前沿的开放权重模型**，为企业组织提供一个起点。
- **GenAIOps 和 DevOps 工具**，它们使 AI 工程师、数据科学家、机器学习（ML）工程师和应用开发人员能够创建、部署和管理 AI 模型、AI 代理以及 AI 驱动的应用。
- **访问模型调优工具，如微调和 RAG 功能**，以便利用私有企业数据自定义模型，并使其适应特定领域的用例。
- **推理运行时**，使您能够提供最佳的性能、吞吐量和延迟。
- **用于 AI 代理的基础组件**，以便在生产环境中管理、治理并确保实施的安全性。
- **计算、存储和网络加速器**，用于加速数据准备、模型自定义和推理任务。
- **基础架构端点**，为 AI 运维的所有阶段提供跨本地、虚拟、边缘以及私有、公共和混合云环境的资源。



本电子书回顾了构建有效 AI 架构的关键注意事项。

推理是 AI 的生产运行时。模型只有在具备 API 并能够提供内容时，才会真正为您所用。这些内容正是通过推理来提供的。

**Chris Wright**

红帽首席技术官<sup>4</sup>

<sup>4</sup> Ron Miller, “[Red Hat’s CTO sees AI as next step for company’s open approach](#)” (红帽首席技术官将 AI 视为公司开放策略的下一步), Fastforward, 2025 年 11 月 11 日。



图 1. AI 架构的组件。

## AI 部署挑战

企业组织面临着选择、构建和交付具有竞争优势的 AI 解决方案的压力。在实施和扩展 AI 部署的过程中，有几项挑战亟待解决：

- **模型成本。** 大规模运行大型模型和推理的成本可能十分高昂。企业组织必须优化模型和推理，以控制计算成本，同时仍保障交付准确且响应迅速的应用。
- **对齐的复杂性。** 模型训练和调优以及创建 RAG 管道的过程既复杂又高度依赖图形处理单元（GPU）算力。企业组织可以简化企业数据的自定义过程，并让业务专家

和开发人员参与其中，从而更快地从实验阶段迈向生产阶段。

- **控制和一致性。** 预打包的 AI 服务会限制对硬件、数据和治理的控制权。选择混合方法，这样您就可以自主选择模型和基础架构，同时保留对数据、生命周期及部署规模的掌控权。

要应对这些挑战，需要一个开放的混合 AI 平台，为跨环境的模型优化、自定义和治理提供一致的工具。

# 容器和容器编排



## 容器

**容器**是一个基本的软件单元，用于打包应用及其所有依赖项。容器可以简化应用的构建流程，并允许在不进行任何更改的情况下跨不同环境部署应用。



### 为什么它们对 AI 至关重要？

AI 工程师和应用开发人员需要访问他们想要的工具和资源，以最大限度地提高生产力。与此同时，IT 运维团队需要确保资源处于最新状态、符合合规要求并以安全的方式使用。

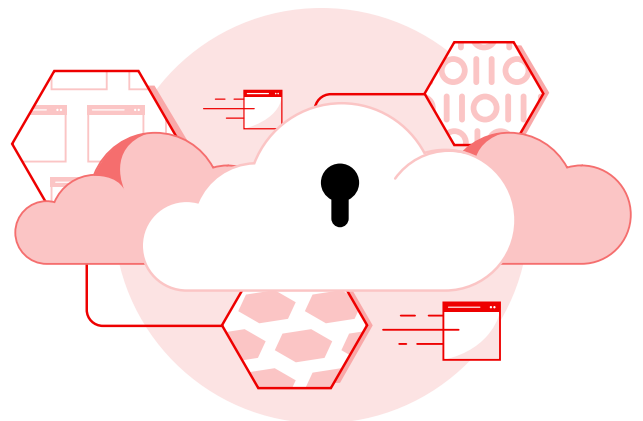
容器通常是部署 LLM 和生成式 AI 驱动的应用的最佳选择，因为它们将模型服务器、依赖项和配置打包成一个可重复的单元，使得生产部署、扩展和更新更易于管理。

借助容器，您可以在混合环境中以一致的方式部署各式各样的 AI 工具。团队可以迭代地修改和共享具有版本控制功能的容器镜像，这些功能可跟踪更改以提高透明度。同时，流程隔离和资源控制提高了防范威胁的能力。



### 最佳实践和建议

寻找一个灵活且高度可用的容器平台，该平台应包含集成的安全防护功能，并简化您在环境中部署、管理和移动容器的方式。选择一个集成了广泛技术的开源平台，以获得更大的灵活性和更多的选择。



# 容器编排

容器编排涉及管理整个环境中容器的创建、部署和生命周期。



## 为什么它对 AI 如此重要？

采用容器后，您需要一种方法来高效地部署、管理和扩展容器。借助容器编排引擎，您能够以一致的方式管理容器的整个生命周期。这些工具通常可以集中访问本地、边缘和云环境中的计算、存储和网络资源。它们还提供统一的工作负载调度、多租户控制和配额执行功能。

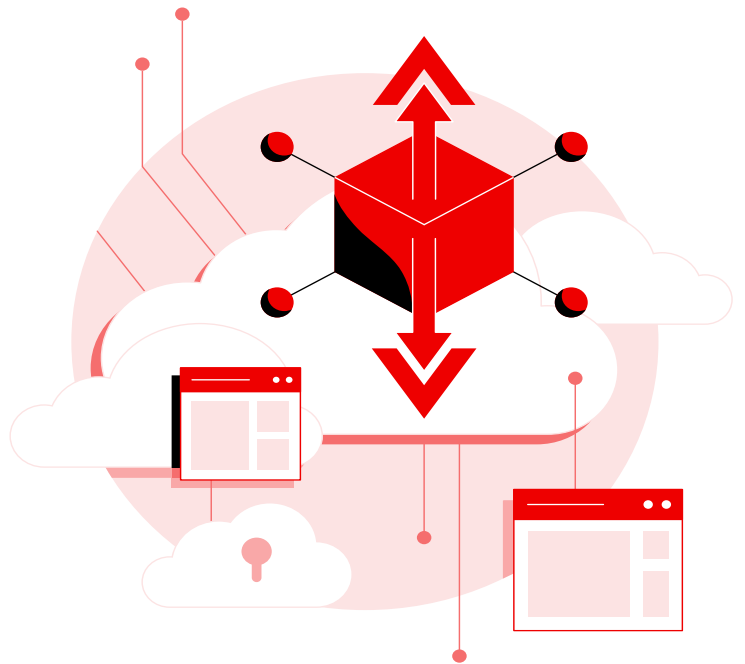


## 最佳实践和建议

选择基于 Kubernetes 的容器编排环境，以领先的开源技术为基础进行构建，并避免专有云锁定。寻找一个能够提供强大的多租户控制、基于角色的访问权限和策略管理功能的平台，以便一致地管理 AI 工作负载。优先考虑那些拥有广泛的 Operator 和集成生态系统的选项，从而标准化在混合环境中部署、扩展和管理 AI 服务的方式。



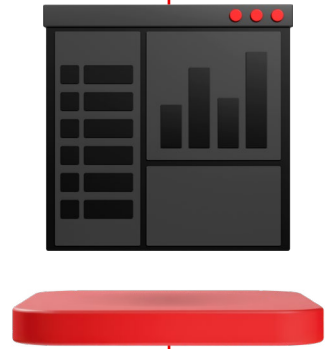
到 2027 年，超过 75% 的 AI 部署预计将采用容器技术作为底层计算环境，而 2024 年这一比例还不足 50%。<sup>5</sup>



<sup>5</sup> Gartner, “Magic Quadrant for Container Management” (容器管理魔力象限), 2024 年 9 月 10 日。

AI 平台注意事项

# 应用管理和 GenAIOps



## AI 工作负载生命周期管理

AI 工作负载生命周期管理专注于如何部署、扩展和管理为 AI 用例提供支持的工具和服务。



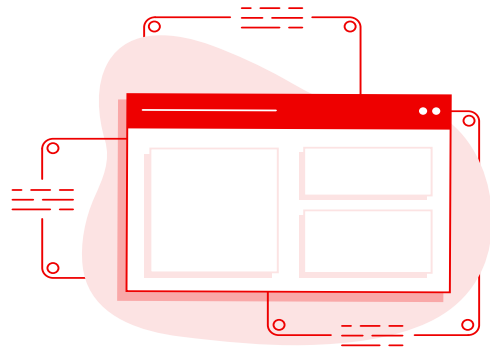
### 为什么它对 AI 如此重要？

AI/ML 环境本身就具有复杂性。AI 工作负载生命周期管理组件（如 Notebook、工作台、管道和模型服务端点）应进行容器化，以便于控制和管理。IT 运维团队可以自动执行常见的生命周期任务，如配置、置备和更新，以提高准确性并减少手动操作。数据科学家、AI 工程师和应用开发人员可以从目录中请求预先批准的 AI 环境，而无需向 IT 部门提交工单。借助自动化，员工能够从重复性任务中解放出来，将更多时间精力投入到价值更高的战略活动中。



### 最佳实践和建议

有效的 AI 工作负载生命周期管理，始于精心筛选的 AI 工作台和 Notebook 镜像。这些镜像包含常用的 AI 和 ML 库，使团队能够从一个安全、受支持的基线环境起步，而非临时搭建的环境。企业组织应提供基于浏览器的 Notebook 环境，并集成 Git，以便团队能够协作进行实验，并持续跟踪代码与模型的变更情况。



# GenAIOps 和 MLOps 实践

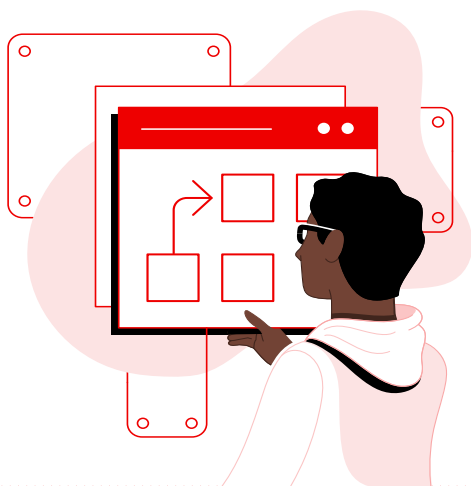
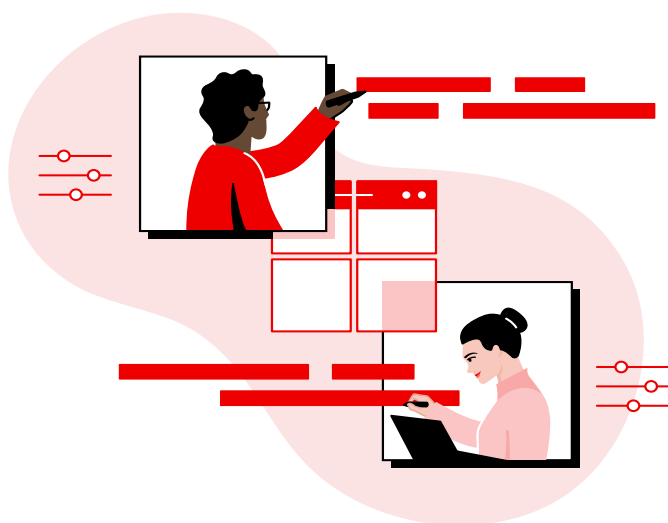
GenAIOps 和 MLOps 实践汇集了大规模实施 AI 所需的工具、平台和流程。



## 为什么它对 AI 如此重要？

企业组织需要快速高效地开发和部署 AI/ML 模型以及使用它们的应用。团队之间的协作对于这些工作的成功至关重要。

与 DevOps 类似，GenAIOps 和 MLOps 方法可促进 AI 工程师、应用开发人员和 IT 运维团队之间的协作，以加快生成式 AI 模型、AI 代理和 AI 赋能应用的创建、训练、部署和管理。自动化通常以持续集成/持续交付（CI/CD）管道的形式出现，使快速、渐进式和迭代式变更成为可能，从而加快模型和应用开发生命周期。



## GenAIOps 和 MLOps 实践

GenAIOps 和 MLOps 不仅涉及技术，人员和流程也发挥着关键作用。将 GenAIOps 和 MLOps 实践应用于整个 AI 生命周期。使用平台和工具中的自动化功能，以及 [Kubeflow](#) 等开源技术来创建 CI/CD 管道和工作流。

# 混合云平台



混合云平台为跨本地、边缘和云环境开发、部署和管理 AI 奠定了坚实基础。它还为您提供了一种方法，能够从初期阶段就针对主权 AI 和私有 AI 进行设计，因此您可以决定哪些工作负载在公共云中运行，哪些保留在您掌控的本地或私有云环境中。



## 为什么它对 AI 如此重要？

AI 模型、代理、软件和应用都需要可扩展的基础架构来进行开发和部署。借助一致的混合云平台，您可以在基础架构的所有部分以相同的方式开发、调优、测试、部署及管理 AI 模型和应用，从而获得更大的灵活性。

它还支持主权 AI 和私有 AI 策略，允许您将敏感数据和模型保留在特定区域，甚至断开连接的环境中，以满足数据驻留、隐私及合规要求，同时在合理的情况下仍然可以连接到公共云服务。自助服务功能可以在保持 IT 控制的前提下，加快资源交付速度。

最后，一致的平台为来自第三方供应商、开源社区以及您使用的任何自定义工具的技术集成奠定了基础。



## 最佳实践和建议

选择一个安全至上的平台，该平台需支持硬件加速、广泛的 AI 和应用开发工具生态系统，以及集成的 GenAI Ops 和运维管理功能。

寻找能够针对数据本地性、模型放置和访问权限提供强大策略控制的平台，以便您能够在本地或私有云中运行主权和私有 AI 工作负载，同时在需要时还能连接到公共云。

若选择开源平台，可获得更多集成机会和更大的灵活性，通过社区驱动的开发促进快速创新，并借助自助服务功能，在保持 IT 控制的同时加快资源交付速度。

公共云和专用本地基础设施的混合组合是最常见的数字基础架构策略。<sup>3</sup>



<sup>3</sup> IDC 白皮书。“AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025”（AI 需求推动本地基础设施部署及与公共云互操作性的需求增长（2025）），文档编号 US53418426，2025 年 10 月。（需要购买）

# 模型自定义和对齐

现代 AI 赋能应用需要能够反映企业组织特定数据、工作流和业务限制的模型。通过将前沿或开放模型与您的专有信息相结合，您可以从获得通用型响应转变为获得准确且具备领域认知的结果。

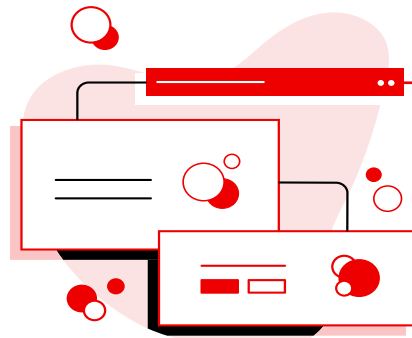
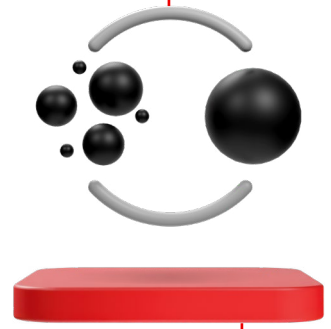


## 为什么它对 AI 如此重要？

生成式 AI 和代理式 AI 依赖于能够理解您的专业术语、数据及实际上下文的模型。

模型对齐通过让模型以您的私有数据为基石，帮助保持其准确性和相关性。它一方面通过降低推理成本并避免不必要的过度配置来提高效率，另一方面能让您将业务逻辑、安全规则及合规要求直接融入到模型行为中，从而加强治理和控制。同时，它还支持可扩展性，为您提供一致的流程，让您能随着数据的演变来更新、重新训练模型并对模型进行版本管理。

模型自定义还支持主权 AI 和私有 AI 策略，使企业组织能够完全在受控环境中训练模型和提供模型服务，以满足数据驻留、隐私和监管要求。



## 最佳实践和建议

采用模块化工作流，根据您的具体需求，从 RAG、微调、提示工程和策略层入手，而非依赖于单一方法。使用开放模型以避免锁定，并保持以透明方式微调、量化和评估模型的能力。让业务专家参与其中，确保模型反映真实的业务上下文和数据准确性。通过应用量化、提炼和高效运行时等技术，尽早优化模型以实现推理，从而控制成本和延迟。此外，通过对版本数据集、训练运行、模型权重和评估指标进行严格管理，以保持可重复性与强有力的治理。

AI 平台注意事项

# 大规模 AI 推理



要在生产环境中运行 AI，需依赖于快速、高效且可靠的推理。模型经过训练或对齐后，便会进入推理阶段。在该阶段，模型会处理新数据、返回预测结果、生成内容或在应用或工作流内部触发相应操作。

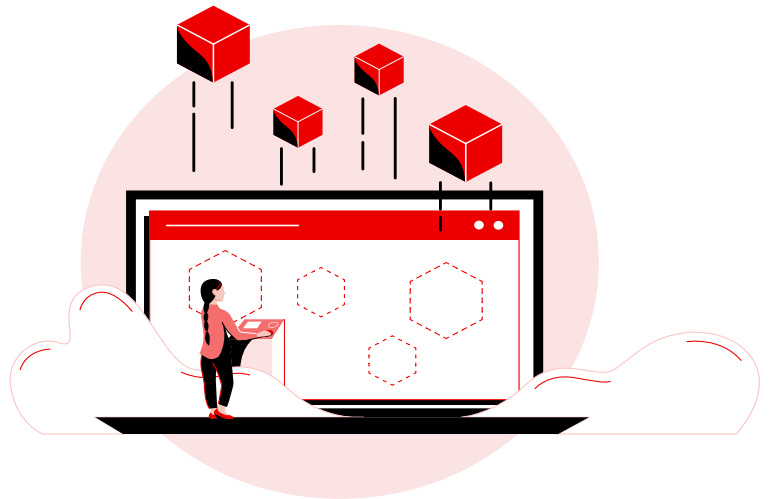
随着企业组织采用生成式 AI 和代理式 AI，推理将成为关键的成本和性能因素，尤其是当应用从单一查询交互转变为由 AI 代理执行的连续、多步骤任务时。



## 为什么它对 AI 如此重要？

推理直接影响用户体验、应用性能和运营成本。生成式 AI 和代理式 AI 工作负载通常需要在许多环境（从数据中心到公共云再到边缘站点）中实现快速响应、并行请求和一致的吞吐量。

高效的推理运行时有助于降低 GPU 和中央处理单元（CPU）成本，改善交互式任务的延迟，并满足调用工具、使用应用程序接口（API）和协调多步骤工作流的 AI 代理的扩展需求。优化推理还支持主权 AI 和私有 AI 策略，允许企业组织在靠近敏感数据的位置（本地或私有云中）运行推理，同时保持可预测的性能。



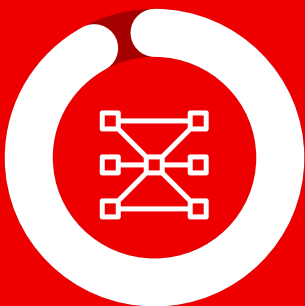
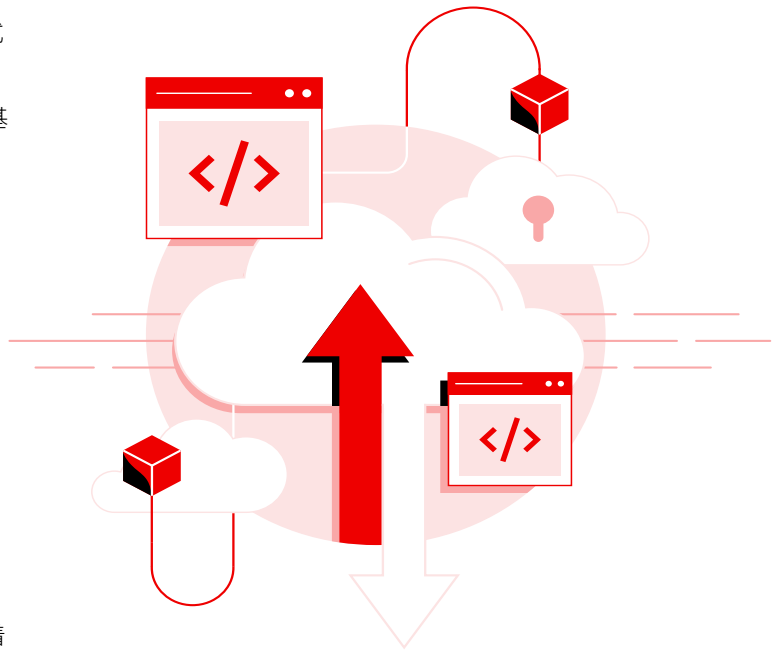


## 最佳实践和建议

选择适合您模型类型和部署环境的优化推理运行时，无论是针对 LLM、多模态模型、预测模型还是代理式工作负载。优先考虑支持动态扩展（包括横向和纵向扩展）的运行时和基础架构，以满足交互式 LLM 和基于代理的推理所带来的不可预测的需求。

在量化、提炼和模型优化等方法中，使用相关技术与具备专业知识的供应商合作，以降低成本并改善延迟。将这些优化措施与广泛采用的技术（如用于高吞吐量 LLM 推理的 vLLM）和新兴的分布式推理框架（如 llm-d）相结合，后者可分解推理过程，使各阶段能够独立扩展。

在容器内部署推理，以打包依赖项，并在混合环境中一致地进行扩展。将推理端点放置在数据和应用所在的位置，以减少数据移动并保持控制，尤其是在主权和私有 AI 场景中。最后，持续监控模型性能，并随着数据分布的变化更新模型版本，以在大规模应用中保持准确性与可靠性。



**90%**

的决策者认为，到 2026 年，AI 将成为其数字基础架构预算和技术选择的重要驱动因素。<sup>3</sup>

<sup>3</sup> IDC 白皮书。“AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025”（AI 需求推动本地基础架构部署及与公共云互操作性的需求增长（2025）），文档编号 US53418426，2025 年 10 月。（需要购买）

# AI 安全防护



AI 系统的行为必须可靠、可预测，并且符合企业组织策略。随着企业从实验阶段迈向生产阶段，AI 安全成为一项核心要求，尤其是在部署能够执行操作而不仅仅是提供建议的生成式 AI、代理式 AI 和自主工作流时。



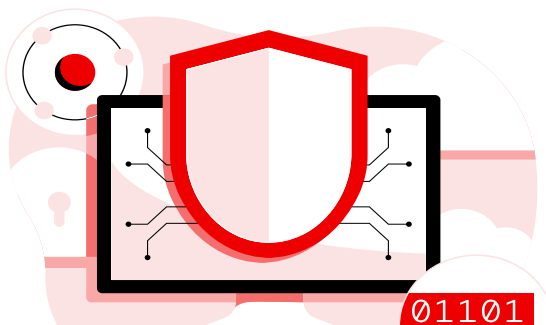
## 为什么它对 AI 如此重要？

安全防护的重点在于确保 AI 模型和代理在规定边界内运行，以符合商业、法律和道德要求。不准确的输出、模型偏移、不安全的数据处理或意外操作，这些都可能导致实际的运营风险。生成式 AI 和代理式系统还引入了新的安全注意事项，如幻觉、未经批准的工具执行、特权升级，以及多步骤任务中的推理不一致问题。强大的安全防护实践有助于您维护信任、保护敏感数据并防止有害或不可逆的操作发生。在受严格监管的行业中，安全控制对于在混合及本地环境中确保合规性与审计就绪状态至关重要。



## 最佳实践和建议

采用分层安全方法，包括针对代理式工作流的基于策略的防护机制、内容过滤器和工具执行控制。定期验证和测试模型，以检测偏移或准确性下降问题。在私有或本地环境中运行敏感工作负载，以保持对数据暴露和模型行为的控制——这与主权 AI 和私有 AI 战略保持一致。使用模型评估框架来监控偏差、稳健性和可靠性。考虑通过相关工具来增强您的模型和数据，这些工具能够将模型和数据存储在符合标准容器（OCI）规范的注册表中，并提供安全的供应链。广泛采用的技术（如用于 LLM 推理的 vLLM）和新兴的分布式技术（如 llm-d），有助于降低成本并扩展 AI 项目部署。最后，对模型、数据集和策略进行版本化和记录，以便您能够在整个 AI 生命周期内跟踪决策并实现一致的治理。



# 构建开放、灵活的 AI 基础



**红帽 AI Enterprise** 是一个集成式 AI 平台，用于跨混合云环境开发和部署高效且经济高效的 AI 模型、代理和应用，并且也是红帽 AI 产品组合的一部分。

它通过提供一个具备企业级能力的即用型开发环境，统一了 AI 模型与应用生命周期，从而提高运营效率、加快交付速度并降低风险。

该平台是一个经过测试且享有支持的完整 AI 堆栈，由红帽 OpenShift 提供支持，可增强互操作性并确保业务连续性。它包含模型调优、高性能推理和代理式 AI 工作流管理等核心功能。这使得平台具备足够的灵活性，能够支持任何模型、使用任何硬件，并在满足数据驻留要求的前提下部署于任何位置。红帽 AI Enterprise 在混合环境可获得支持，因此团队可以放心地规划容量、GPU 以及未来的 AI 项目。

红帽 AI Enterprise 包含来自开源 llm-d 项目的技术，该项目由红帽与 IBM、NVIDIA、Google、AMD 等合作伙伴共同发起。llm-d 通过分离推理的预填充和解码阶段，使每个阶段都能够以不同的方式进行扩展，从而提高成本效率。其具备推理感知能力的负载均衡器能够基于令牌队列来路由请求，从而缩短响应时间，并在某些情况下将预填充工作负载引导至 CPU。



## 缩短价值实现时间。

利用预配置工具、自动化部署和内置可观测性，在您选择的基础架构上部署企业就绪型 AI 堆栈。这意味着开发人员和 AI 工程师可以专注于构建和交付由云原生 AI 驱动的代理式应用。



## 提高运维效率。

简化和自动化工作流，从代码提交到建立 AI 管道工作流，再到模型部署。这意味着 IT 运维能够通过智能资源分配和集成化生命周期管理，实现稳定可靠的性能表现，并从现有基础架构中获取更高价值。



## 缓解风险

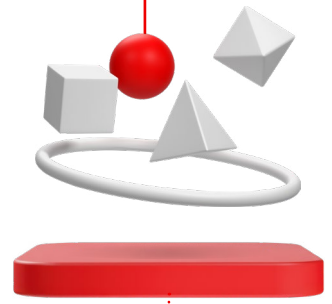
借助经过测试且享有全面支持的集成式 AI 堆栈，降低企业采用 AI 的风险，该堆栈可增强任何模型、任何硬件及混合云环境之间的互操作性。以此为基础来满足数据驻留和监管要求，从而自信地扩展 AI。





图 2. 集成式 AI 平台的组成部分。

# 利用经过认证的 AI/ML 合作伙伴生态 系统获得更多选择 和更大的灵活性



AI 工具和技术格局持续快速演变，这使得在紧跟技术发展步伐的同时，于 IT 环境中保持稳定性和可靠性变得颇具挑战。

通过与英伟达、AMD、英特尔及其他 AI 技术合作伙伴携手合作，红帽 AI Enterprise 提供了一个可在混合云中扩展的端到端企业级 AI 平台，助力实现更快的部署、更高的效率，并提供混合云支持。红帽的验证和认证计划可确保硬件得到充分利用，而优化的工作负载管理则保证了 GPU 的高效使用，从而为客户最大限度地提升性能与价值。

随着红帽加入 [Hugging Face 生态系统](#) 和 [模型上下文协议 \(MCP\)](#) 服务器目录，客户可以访问一个不断扩大的经验证模型和预集成工具库，这些模型和工具能够与红帽 AI Enterprise 一致地运行。与此同时，与多个加速器供应商建立合作伙伴关系，有助于企业组织在混合环境中充分利用 GPU 和专用硬件。您可以放心地选择最契合您需求的合作伙伴、模型、工具和技术，因为它们将可靠地协同工作，并以专家服务、支持和培训为后盾，助您成功构建和扩展 AI 工作流。



# 合作伙伴亮点



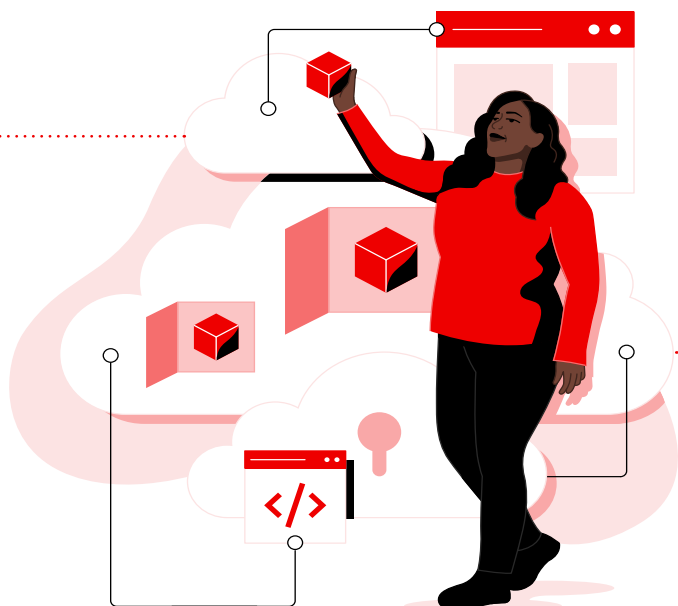
[NVIDIA](#) 是 AI 技术领域的全球领导者，为 AI 和加速计算提供了领先的 GPU 架构和平台，在训练和部署大型 AI 模型方面实现了卓越的性能和能效。英伟达与红帽之间建立了战略合作伙伴关系，专注于为 AI 提供优化的云原生基础架构和软件。通过与红帽 AI Enterprise 相结合，客户将获得一整套经过优化的云原生 AI 和数据分析软件，以及运行这些软件所需的硬件。红帽 AI Enterprise 和 NVIDIA DGX 系统共同为 AI 基础架构提供 IT 可管理性。[NVIDIA GPU Operator](#) 可自动管理置备 GPU 所需的所有 NVIDIA 软件组件。



[AMD](#) 与红帽携手合作，旨在扩大混合云环境中 AI 与虚拟化的选择范围。此次合作将 AMD EPYC 处理器和 AMD Instinct 加速器引入红帽 AI Enterprise，使企业组织能够更加灵活地在具备成本效益的高性能硬件上运行 AI 训练、调优和推理工作负载。联合工程和认证工作可确保由 AMD 提供支持的系统与红帽 AI Enterprise 一致地运行，包括支持优化的容器镜像、Kubernetes Operator，以及对 GPU 加速 AI 管道的性能验证。



[英特尔](#)与红帽携手提供软件定义基础架构和行业标准平台，以提高数据中心的敏捷性和灵活性。英特尔发行的 [OpenVINO 工具包](#)可优化 AI 模型并将其转换为高性能推理引擎，这些引擎可以自动扩展到红帽 AI Enterprise 上的数千个节点。



# 成功案例



## 土耳其航空公司

土耳其航空利用红帽 AI 实现运营现代化，并在整个航空领域引领 AI 驱动的创新。通过在开放、可扩展的 AI 平台上实现标准化，该航空公司加快了模型开发速度，优化了乘客服务，并简化了运营决策过程。这一案例展现出混合 AI 如何驱动全球最大的航空公司网络之一实现转型升级。

[了解更多信息](#)

## Denizbank

Denizbank 利用红帽 AI 加速其数字银行生态系统中的 AI 创新。通过使用开放、可扩展的平台对其 AI 基础架构进行现代化改造，该银行加快了实验速度，提高了模型可靠性，并提供了更智能的客户体验。这一案例展现出混合 AI 如何帮助金融机构在保持严格安全态势与治理的同时，加快前进步伐。

[了解更多信息](#)

## AGESIC

AGESIC 是乌拉圭的数字政府机构，它利用红帽 AI 在 180 多个公共实体中实现 AI 的标准化与规模化应用。该混合 AI 平台支持 MLOps 实践，增强安全性，并帮助各团队构建、部署和治理各类 AI 应用，从而为公民提供更好的服务。

[了解更多信息](#)



# 准备好从您的数据中 获得更大价值了吗？



AI 正在改变业务的方方面面。红帽能帮助您打造一个生产就绪型 AI 环境，从而加快智能应用的开发和交付，以支持您的业务目标。

[进一步了解红帽 AI Enterprise 如何助您构建统一的 AI 平台。](#)

