依托红帽 AI Enterprise 平台，打造集成式 AI 体系

AI 的快速崛起

AI 已得到广泛采用，并取得了以下成果：

- 55% 的企业组织正在使用生成式 AI。[1]
- 超过 50% 的企业组织正在进行代理式 AI 概念验证或实施特定用例。[2]
- 预计到 2026 年底，30% 的 AI 赋能应用将采用代理式 AI。[2]

AI 机遇不断演进，复杂性也随之攀升

AI 为企业提供了重塑运营模式和竞争优势的契机，并已迅速成为当今商业格局中的关键要素。然而，尽管 AI 的采用率不断上升，许多企业仍面临诸多挑战，阻碍了其充分发挥价值。

这些挑战包括不断上升的模型成本、复杂的定制需求、严格的部署限制，以及跟上不断加快的创新步伐等。

随着企业将重心转向生成式 AI 创新、代理式 AI 工作流和主权 AI，应对这些挑战的难度也与日俱增。因此，企业亟需展开向 AI 平台的战略迁移，以帮助降低推理成本，简化扩展与监控流程，构建本地主权 AI，并迅速适应变化。

利用企业级平台和工具，化解 AI 复杂性

无论企业是刚刚开启 AI 之旅，还是正在整个企业组织内规模化推广成熟的 AI 计划，红帽® AI 都能助力加速 AI 创新，并降低在混合云和多云环境中开发和交付 AI 解决方案的运维成本。

红帽 AI 提供了经济高效的解决方案，可以帮助实现模型优化和高效推理，简化与私有数据的集成，并在可扩展且灵活的平台上加快交付代理式 AI。它使企业组织能够大规模管理和监控预测性 AI 与生成式 AI 模型的整个生命周期，从单服务器部署到高度扩展的分布式平台皆可涵盖。

红帽 AI 以成熟可靠的开源技术为基础，为众多领先的开源 AI 模型提供支持，帮助企业加速创新并普及对宝贵工具和技术的使用。

此外，红帽 AI 还拥有 AI 合作伙伴生态系统，该系统提供经过测试和认证的产品与服务，专注于跨基础架构的性能、稳定性和加速器支持，助力企业借助 AI 解决关键业务和技术挑战。

红帽官方微博　红帽官方微信

cn.redhat.com

1 IDC Tech Buyer Presentation。“Worldwide Generative AI Industry Use Case Early Adoption Trends, 2025:Executive Summary”（2025 年全球生成式 AI 行业用例早期采用趋势：内容摘要）。文档编号 US53280825，2025 年 4 月。（需要购买）

2 IDC Survey。“Agentic AI Impact on Digital Infrastructure Strategies”（代理式 AI 对数字基础架构策略的影响）。文档编号 US53418526，2025 年 10 月。（需要购买）

红帽 AI 具备以下优势：

- 完整的 AI 模型生命周期监控与管理。
- 支持在任何硬件加速器上灵活使用任何 AI 模型。
- 经过优化和验证的 AI 模型集合。
- 跨混合云或多云环境进行灵活且经济高效的推理。
- 安全可信地与企业私有数据资产集成。
- 简化代理式 AI 工作流的交付。
- 按需扩展，高效利用资源。

红帽 AI Enterprise 是一款统一 AI 平台，作为红帽 AI 产品组合的一部分提供，可帮助企业解决当下采用 AI 所面临的挑战。

选择红帽 AI Enterprise 作为 AI 平台的六大理由

红帽 AI Enterprise 可提供企业所需的所有工具和功能，在其 AI 之旅的任一阶段提供支持。它使企业能够在统一的 AI 平台上，跨混合云和多云环境开发和部署高效且经济的 AI 模型、代理及应用。

该平台还为企业提供了随时可用的开发环境以及关键 AI 功能，包括模型调优、高性能推理、代理式 AI 工作流管理；并具备支持任何模型、使用任何硬件以及在任何位置部署的灵活性，以满足不断变化的数据驻留要求。

以灵活高效的推理，简化 AI 部署

随着 AI 创新持续加速，部署的复杂性也随之提升，这可能会导致资源利用率低下、运维成本增加以及成果质量下降。

红帽 AI Enterprise 可在本地、混合云、多云或网络边缘等不同环境中提供优化的模型推理，助力企业简化模型部署，同时减少资源使用、降低推理成本并保持准确性。

这一优势得益于由 vLLM 提供支持的优化运行时。vLLM 是一款推理服务器，可通过更高效地利用 GPU 内存来优化吞吐量和延迟，从而加速依托生成式 AI 的应用的输出。此外，该平台还采用先进的压缩技术，通过减小模型规模和降低计算需求，进一步优化成本。

对于准备进行扩展的企业组织，例如从几个模型扩展到数十个以及从数十个用户扩展到数百个，该平台提供了使用 llm-d 的选项。这项技术基于 vLLM 的核心创新成果而构建，提供了一个分布式推理框架，可在生成式 AI 模型的大规模应用中确保实现经济高效、可预测的性能，并且随着处理量的增加，性能还会进一步提升。

利用企业私有数据，提高 AI 模型准确性

尽管公开可用的生成式 AI 模型功能强大，但企业仍需要基于自身独有数据训练的 AI 模型，以满足特定领域的用例需求。

红帽 AI Enterprise 可为企业提供简化且一致的体验，以支持高效的模型定制，并帮助 AI 工程师和数据科学家提高模型响应的准确性和相关性。该平台支持将模型与企业组织自有数据相连接，以支持检索增强生成（RAG）用例。

该平台还支持微调、持续学习和强化学习，并为文档处理和解析、合成数据生成以及模型评估和任务提供灵活的模块化方法。

借助代理式 AI 工作流，加速创新步伐

AI 技术在相对较短的时间内实现了迅速发展。从预测性 AI 开辟了众多新的用例，到生成式 AI 模型的普及，再到如今聚焦于代理式 AI 工作流以推动新一轮演进，这将进一步扩展 AI 战略，为企业带来更多商业价值。

红帽 AI Enterprise 提供了一个敏捷、稳定的平台，助力企业利用代理式 AI 成功实现价值。为实现这一目标，该平台提供了统一的应用编程接口（API）层、开箱即用的组件（可简化企业用例下代理式 AI 工作流的搭建）、专属用户体验，以及一个灵活且可扩展的基础，用以支持代理式 AI 系统的部署与管理。

该平台还支持模型上下文协议（MCP），这是代理式部署的关键组件，充当各种工具和功能与大语言模型（LLM）之间的标准化转换器。

在混合云中扩展 AI，并保持灵活性和一致性

许多行业的企业组织已经在初始测试用例中验证了 AI 的价值。然而，其中许多企业仍然难以将这一价值扩展至整个企业，同时兼顾成本效益，并确保监管合规性、隐私性和安全性。当涉及混合云、多云等多样化 IT 环境或网络边缘的 AI 部署时，这一挑战尤为突出。

红帽 AI Enterprise 可为企业提供所需的控制力与一致性，支持在任意硬件或 IT 环境中，根据企业组织的 AI 工作负载战略，灵活训练、调优、部署、管理及扩展预测性 AI 和生成式 AI 模型。

该平台具备高度灵活性，支持在硬件、原始设备制造商（OEM）、云提供商与数据中心的任意组合环境中运行任意 AI 模型。结合全面的可观测性与监控能力，企业可在规模化扩展过程中优化性能、控制成本、强化治理，并始终聚焦安全防护，以满足不断演变的业务需求。

通过本地 AI 部署，维护数据主权

世界各地政府不断出台关于 AI 使用以及数据存储和处理方式的法规。这促使许多企业探索如何构建本地 AI 部署，以维护主权和隐私。

通过在本地存储和处理敏感的私有数据，企业不仅可以确保遵循许多不断演变的法规，还可以通过缩小攻击面和加强安全防护来保障数据隐私。

红帽提供全面、分层的 AI 安全防护方法，让企业组织能够自信地拥抱 AI 创新。红帽 AI Enterprise 基于值得信赖的开放混合云提供这一安全防护方法，以帮助应对 AI 生命周期中独特的安全防护挑战。

该平台支持任意模型、任意硬件，以及在所有云与边缘环境（包括隔离环境）中实现安全至上的 GPU 共享，助力企业实现真正的主权控制。红帽与主权云提供商密切合作，使企业组织能够在注重安全的私有云环境中部署 AI。

红帽 AI Enterprise 还内置相关工具，通过模型可解释性、公正性、输出策略管控实现可审计的信任机制，并借助可再现管道与审计就绪性，实现整个生命周期的可追溯性。

在值得信赖的 Kubernetes 平台上构建 AI

红帽 AI Enterprise 基于可靠的 Kubernetes 基础架构，以集成式、随时可用的端到端 AI 技术堆栈的形式，提供这一整套完备的 AI 功能。

这为已使用 Kubernetes 和容器化的企业，在本地数据中心、混合云或多云环境以及边缘 AI 部署中提供了熟悉且一致的体验。企业可基于单一集中化平台，使用团队已熟练掌握的工具、框架和技能，构建、训练、部署及管理现代化的云原生 AI 应用。

进一步了解红帽 AI Enterprise 的价值

详细了解红帽 AI Enterprise 如何凭借统一平台支持 AI 之旅各个阶段，助力企业在混合云中部署高效且经济的 AI。

销售及技术支持

800 810 2100
400 890 2100

红帽北京办公地址

北京市朝阳区东大桥路 9 号侨福芳草地大厦 A 座 8 层　邮编: 100020
8610 6533 9300

红帽官方微博　红帽官方微信