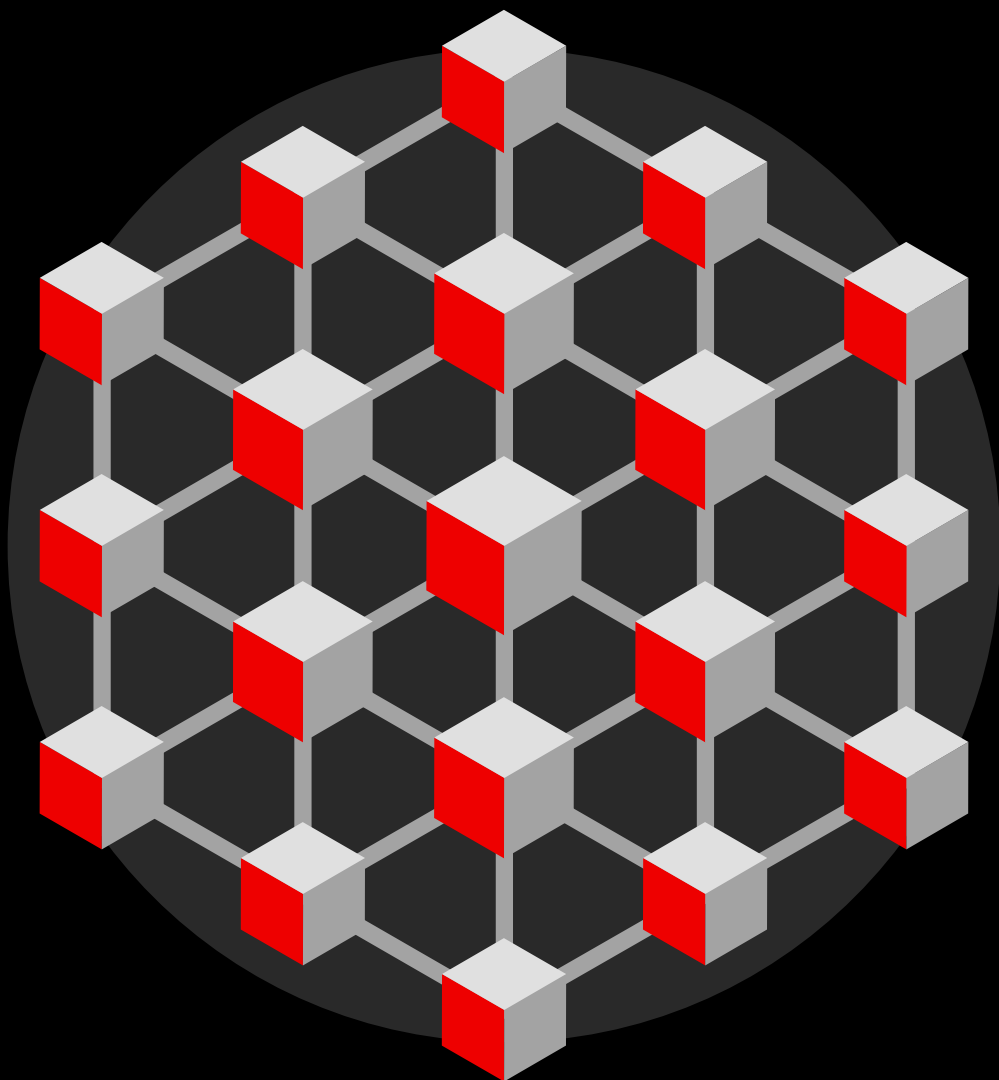


利用**开源模型**， 释放 AI 创新潜力



Red Hat

目录

1

简化企业级生成式 AI 部署

2

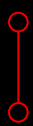
利用适用于特定任务的小语言模型加速创新

3

利用开源生成式 AI 解决方案获得掌控力

4

利用红帽企业 Linux AI 简化 AI 开发流程

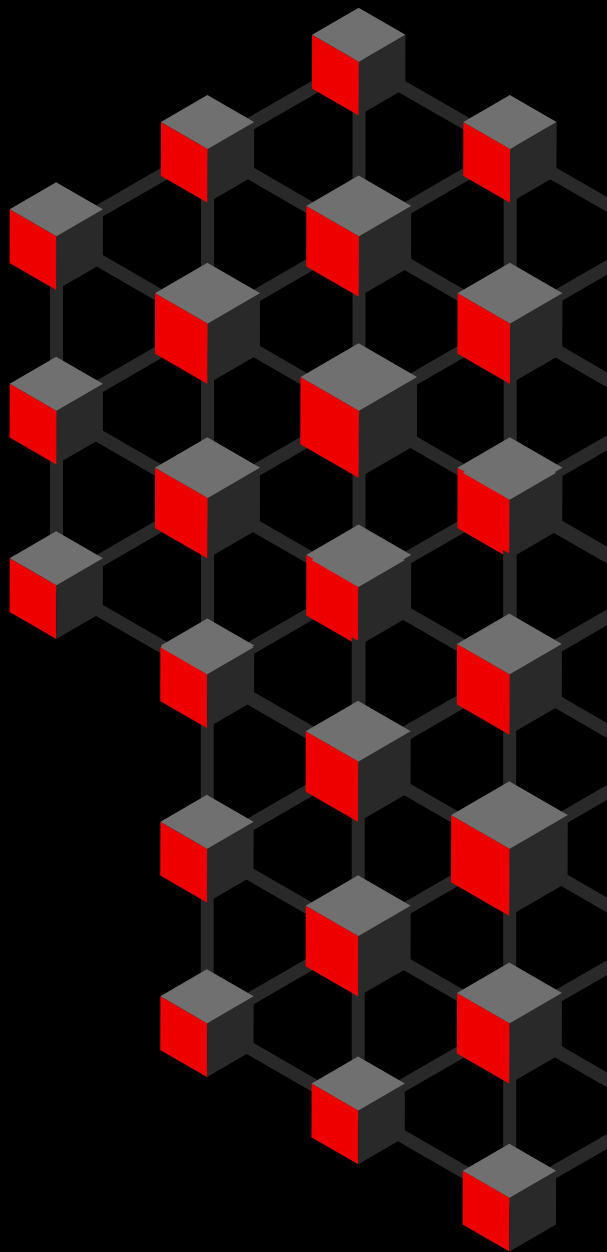


4.1 针对企业 AI 采用开源 Granite 模型

4.2 利用 InstructLab 实现 AI 开发大众化

5

携手红帽加速 AI 之旅



简化企业级生成式 AI 部署

在各行各业，生成式人工智能（生成式 AI）不断革新着企业组织管理运维、制定决策和提供创新型解决方案的方式。

从个性化客户体验、自动化内容创建到加强决策支持，生成式 AI 可帮助企业简化流程、提高生产力并探索新的可能性。通过投资创新型 AI 解决方案，企业组织可以增强驾驭复杂性、应对不断变化的需求以及实现战略目标的能力。

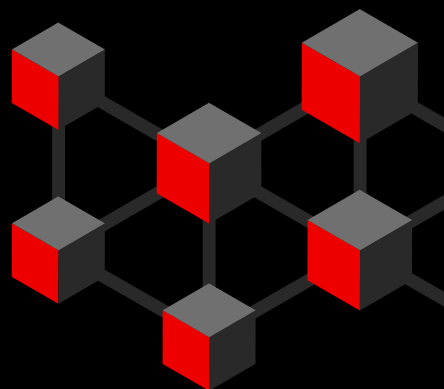
然而，对许多企业组织来说，在充分发挥生成式 AI 技术潜力的过程中面临着诸多挑战。专有大语言模型（LLM）许可费用高昂，训练和推理过程都需要耗费大量硬件资源，因此在企业 IT 环境中部署和维护这些模型的成本非常高。如果开发人员和企业组织主题专家完全了解应用和业务需求，但缺乏必要的专业数据科学知识，那么针对特定客户和行业用例来训练和调整大语言模型可能会十分困难。使用在复杂的混合云环境（包括私有数据中心、公共云资源和边缘部署）中存储的专有机密数据，难以有效地训练生成式 AI 模型，尤其是在监管要求限制数据在不同位置之间移动的情况下。

红帽提供了一种简化的方法，使企业组织范围内的开发人员和领域专家都能使用生成式 AI。借助红帽® AI 解决方案，您可以在生产级服务器部署中高效训练和运行生成式 AI 模型，让团队能够在 AI 旅程中快速前进，无需应对不必要的复杂性。



的企业组织认为，生成式 AI 将在未来 18 个月内颠覆其业务¹。

利用适用于特定任务的小语言模型加速创新



大语言模型具备自然语言理解和生成功能，但资源密集型训练和推理过程需要耗费大量计算成本，这会使该技术在许多业务应用中难以落地。所幸，对大多数企业组织而言，并不需要如此广泛的功能，针对性更强、适用于特定任务的**小语言模型（SLM）**往往更能满足需求。

小语言模型规模较小，所需的计算资源、数据和能耗也更少，是一种高效、经济的生成式 AI 模型，适用于许多企业应用。这类模型训练速度更快、微调流程更精简，并且能提供更加可持续的 AI 开发方法。小语言模型通过简化部署和提高基于 AI 的应用的性能，可帮助降低成本，同时提供量身定制的创新解决方案，以达成您的业务目标。

小语言模型为企业组织提供了多项优势。小语言模型规模较小且对数据要求较低，因此企业组织可以针对特定任务或领域快速高效地定制，从而生成更准确且更具成本效益的生成式 AI 模型。通过在训练过程中将企业组织知识和专业技能直接嵌入到小语言模型参数中，可提高模型相关性、减少重复训练需求，并缩短基于 AI 的关键应用和服务的开发时间。小语言模型可以在本地企业 IT 或私有云环境中高效部署，因而能够更轻松地保护敏感的专有训练数据，同时保持对访问权限的控制、简化监管合规流程，并降低对外部提供商的依赖。在资源有限的环境和边缘部署中，小语言模型允许实时应用直接在用户设备上运行，从而简化开发流程，并降低对外部云基础架构的依赖。



观看本视频，进一步了解小语言模型的优势。

利用开源生成式 AI 解决方案 获得掌控力

使用商业 AI 解决方案常常会引发人们对可见性、安全性、隐私保护和可靠性的担忧，而训练数据来源不确定及响应准确性问题则可能会增加企业组织的法律风险。为应对这些挑战，大语言模型正日益走向开源化，通过提供自由度、选择权和透明度，帮助用户更深入地了解、定制并信任所使用的模型。

开源大语言模型和小语言模型包含已获得开源许可的软件组件和模型权重（由可信赖的提供商预训练），让您能够以透明、可靠的方式为基础模型和训练数据集贡献数据和专业知识。这些模型还有助于避免供应商锁定，使您能够保留对 AI 解决方案的完全控制权，无需依赖专有系统。由于值得信赖的提供商会披露用于训练开源模型的数据来源，因此您可以在添加自己的专有和机密数据之前，对模型质量进行评估，确保其中不包含有害或有偏见的数据。此外，开放式治理实践有助于在模型开发过程中识别并消除偏见，增强 AI 系统的可信度和公平性。

最后，借助根据 Apache 2.0 许可证发布的开源大语言模型，您可以灵活地定制和控制模型，这确保您能够调整创新型 AI 解决方案，以满足业务需求。

开源在 AI 领域的强大威力

开源方法逐渐成为在生成式 AI 技术领域建立信任的重要因素。进一步了解开源在 AI 领域的影响：

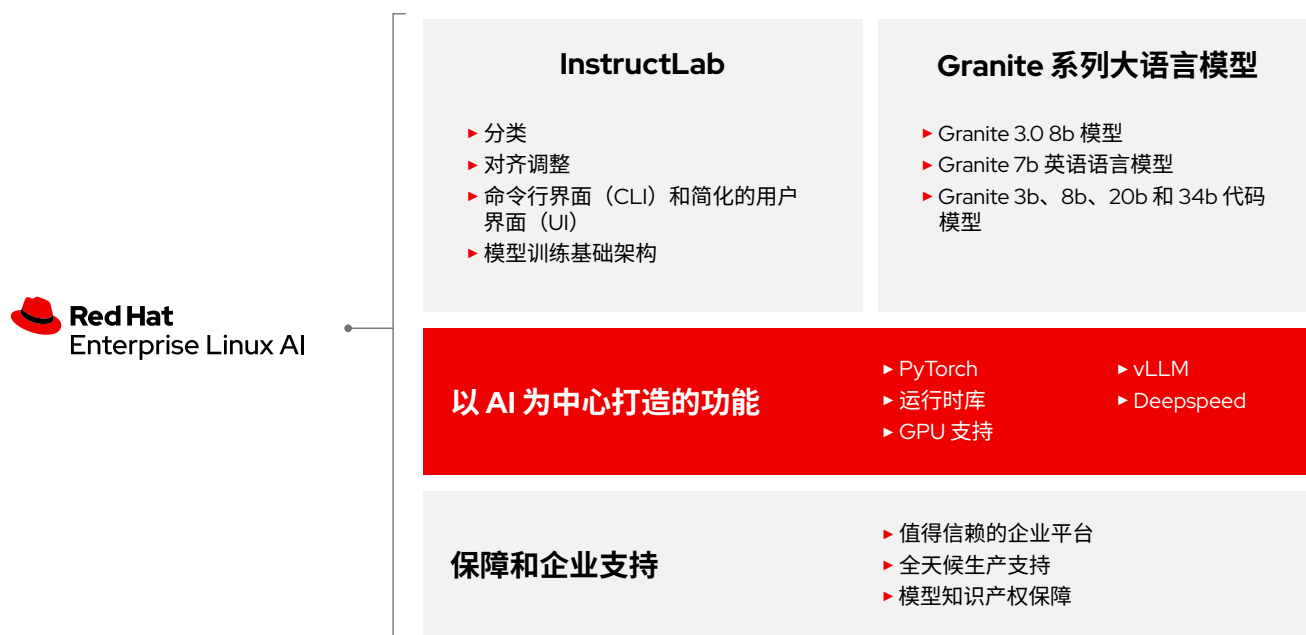
- ▶ [阅读这篇博客文章](#)，了解开源社区如何推动 AI 技术发展。
- ▶ [阅读这篇博客文章](#)，了解为什么越来越多的企业组织对开源 AI 技术越来越信任和有信心。
- ▶ [观看本视频](#)，了解如何为开源 AI 计划做出贡献。

利用红帽企业 Linux AI 简化 AI 开发流程

对于许多企业组织来说，提供既具有成本效益又能满足业务需求的 AI 解决方案可能会是一个难题。红帽企业 Linux® AI 提供了一个一致且稳定的基础模型平台，可加快和简化创新的企业就绪型生成式 AI 解决方案的开发和部署流程。

红帽企业 Linux AI 提供用于训练、测试和运行适用于特定任务的小语言模型的高级功能，让您能够打造经济高效的定制化 AI 应用和服务。作为核心平台组件，Granite 系列获得开源许可的生成式 AI 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。这些高效的精简模型可在保障性能的同时降低运维成本，显著缩短部署时间。此外，InstructLab 模型对齐工具可简化微调流程，通过将模型与企业组织数据对齐，并扩大对社区开发的生成式 AI 模型的访问权限，让开发人员和领域专家都能轻松使用 AI 技术。

红帽企业 Linux AI 以可启动镜像形式交付，内置 PyTorch 等热门 AI 库以及来自英伟达、英特尔和 AMD 的硬件优化加速器，可简化关键 AI 技术的集成，同时简化生产服务器上的模型训练与推理过程，让您能够高效地启动生成式 AI 项目。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。



针对企业 AI 采用开源 Granite 模型

要想打造成功的 AI 应用和服务，需要能够平衡透明度、成本和性能，同时又能满足企业应用需求的模型。由 IBM 开发的 Granite 系列开源生成式 AI 模型，可满足企业应用和真实业务场景的需求。这些模型针对成本和性能进行了优化，支持涉及语言和代码的各种生成式 AI 用例。开源 Granite 模型系列包括 Granite 3.0 8b 模型、Granite 7b 英语语言模型以及 Granite 3b、8b、20b 和 34b 代码模型。



观看本视频，了解有关开源 Granite 模型系列的更多信息。

Granite 模型根据 Apache 许可证 2.0 发布，提供完全开源的可访问性。这些基础 AI 模型是独立构建的，并非源自专有或非开源模型。此外，IBM 还披露了用于训练的数据集和方法，为您的关键 AI 解决方案提供更高的透明度与可信保障。阅读 [Granite 3.0 语言模型技术白皮书](#)，深入了解训练数据和实践，见证 IBM 对开放性与责任担当的坚定承诺。

由红帽提供全面保障的开源 Granite 语言和代码模型，为开发和部署量身定制的 AI 应用提供了企业就绪型解决方案。通过将 IBM 创新模型与红帽平台的可靠性和支持服务相结合，您可以自信地采用开源生成式 AI 解决方案来应对业务挑战，同时降低整个企业组织的风险。

利用 InstructLab 实现 AI 开发大众化

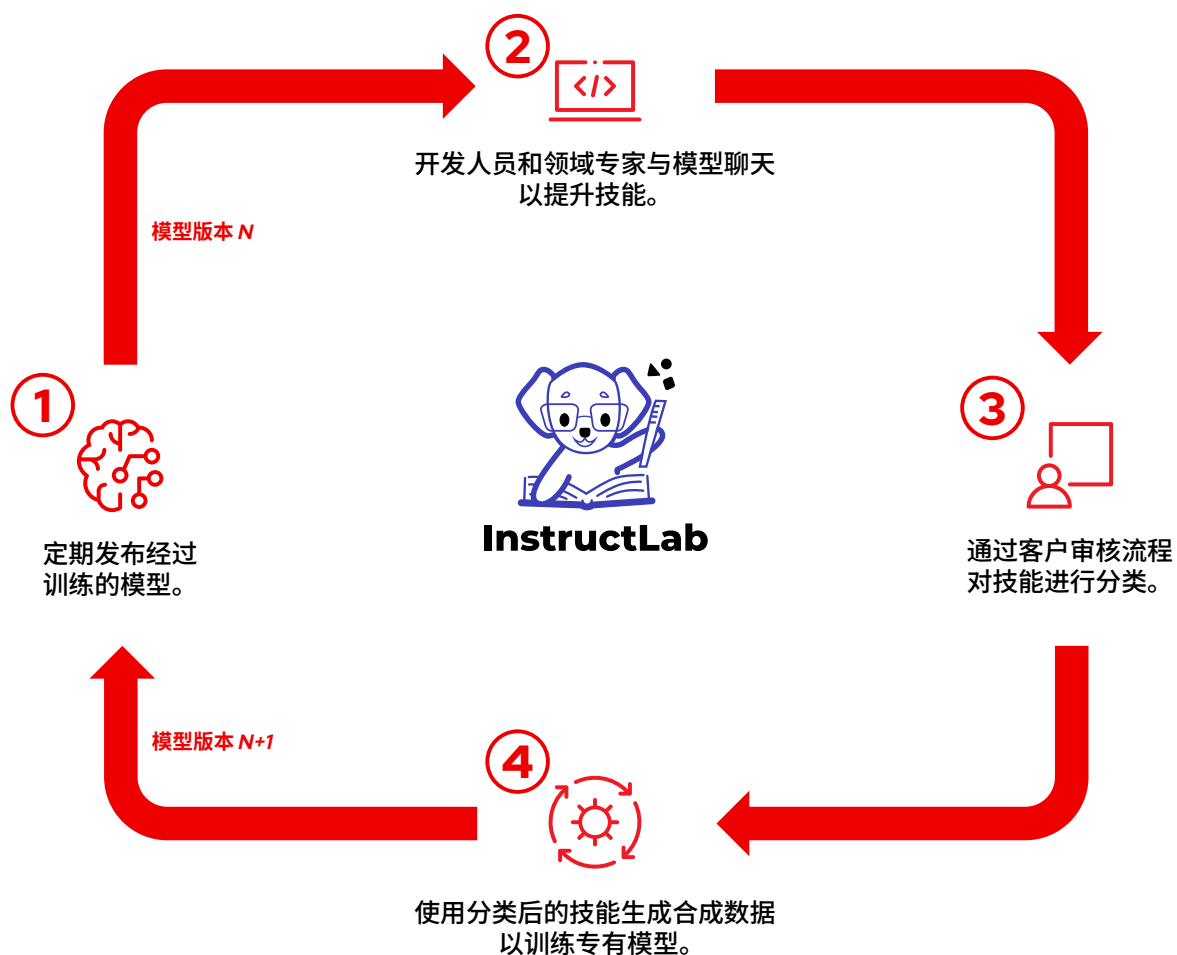
根据业务需求调整生成式 AI 模型可能是一项艰巨且耗时的任务，尤其是在专业知识和资源有限的情况下。借助 InstructLab 模型对齐工具，您可以利用专业知识和技能定制生成式 AI 模型。InstructLab 基于 [聊天机器人大规模对齐 \(LAB\)](#) 技术，采用分类驱动的方法，利用高质量的合成数据生成和训练模型。因此，与以往重新训练模型的方式相比，您可以使用更少的人工生成信息和计算资源来增强 AI 模型。

InstructLab 简化了训练和微调过程，让更多的贡献者（包括没有丰富数据科学经验的开发人员和领域专家）能够参与生成式 AI 模型开发。通过让主题专家直接参与构建创新型 AI 解决方案，InstructLab 促进了技术团队与业务利益相关者之间的协作，确保 AI 应用和服务切合实际需求。通过使 AI 模型调整大众化，可缩短价值实现时间，帮助整个企业组织完善模型，更有效地实现业务目标。因此，您可以根据运维和战略目标调整生成式 AI 模型，提高关键 AI 解决方案的准确性和相关性。通过以经济高效的方式整合企业业务知识，InstructLab 可以帮助您在 AI 投资方面更快取得成果并获得更高的回报。



观看此视频，了解如何利用 InstructLab 训练生成式 AI 模型。

InstructLab 支持以迭代方式微调生成式 AI 模型。它鼓励协作，允许团队成员贡献自己的专业知识，同时由主题专家验证和批准输入内容。这一流程有助于确保您的模型在迭代过程中始终保持透明、相关且无偏见。



携手红帽加速 AI 之旅

为 AI 创新筑牢可信赖的根基。

无论您处于 AI 之旅的哪个阶段，红帽企业 Linux AI 都能提供成功所需的一致性、可靠性和创新性。凭借值得信赖的开源运维基础和集成的 AI 技术，您能够更轻松地启动 AI 项目，并在更短的时间内将创意从概念推进至生产阶段。

了解有关红帽企业 Linux AI 的更多信息。

免费试用红帽企业 Linux AI 60 天

免费获取为期 60 天的自助服务式订阅和资源，了解企业组织如何针对企业应用开发、训练、测试和运行 Granite 系列生成式 AI 模型。

在专家指导下更快上手使用

与我们的专家合作，规划和构建能带来真正业务成果的 AI 解决方案。无论您是刚刚开始使用生成式 AI，还是想扩展 AI 功能，红帽服务都能为您提供帮助。