使用开源小语言模型的四个原因

小语言模型正在重塑企业 AI 战略

专有大语言模型（LLM）在通用应用领域表现出色，但它们并不总是企业人工智能（AI）解决方案的最佳选择。这些模型对计算能力要求极高、决策过程不透明且授权成本高昂，往往会限制灵活性并增加运维复杂性。相比之下，小语言模型（SLM），尤其是基于开源原则构建的模型，为寻求开发定制化 AI 解决方案、保持对数据的掌控并有效控制成本的企业组织提供了另一种选择。

开源 SLM 具有以下四大优势，这亦是其可成为您下一个 AI 项目理想之选的四个原因。

1 获取社区创新成果

开源 SLM 兼具灵活性、协作性与创新性，为构建高度适配且专业的 AI 应用提供了坚实基础。开源 AI 项目既提供软件组件，又开放预训练模型权重，让您能够与全球开发人员及研究人员社区展开合作，持续优化和改进生成式 AI（Gen AI）技术。在这种共享创新模式下，您可以运用现代化的先进工具，并对其进行定制以满足企业 AI 解决方案的技术需求。

借助 IBM Granite 系列生成式 AI 模型等开源 SLM，您可以直接将专业知识和领域专长融入基础模型。相较于被动等待专有 LLM 的更新，您可以主动定制开源 SLM，从而显著提升其在 AI 应用中的适用性和性能表现。这种交互式方法不仅有助于加快模型迭代周期，还能使模型紧跟不断变化的业务需求。

开源 SLM 为动态环境部署提供了至关重要的灵活性，可跨本地数据中心和公共云基础架构轻松部署。您还可以完全掌控模型，从而针对从高合规性环境到实时 AI 处理的各种部署场景进行优化。此外，开源 SLM 有助于您控制 AI 技术栈，确保随着技术和业务需求的变化，您的创新型 AI 解决方案能够保持适应性和可扩展性。

2 掌控训练数据

与专有替代方案相比，开源 SLM 具备更高的透明度。由于值得信赖的供应商会披露用于预训练这些模型的数据，您可以全面评估模型质量，并确认其中不含任何有害或带有偏见的信息。这种透明度使您能够在调整和部署模型时做出明智的决策。因此，在整合自身的专有保密数据之前，您可以确保 AI 解决方案符合伦理标准且契合业务目标。

此外，您可以在企业 IT 环境中的本地数据中心和私有云资源中部署 SLM，因此能够完全掌控训练数据。这种控制权对于处理高度机密或受监管数据的企业组织而言至关重要，因为它能够确保专有信息不会泄露给外部提供商。而且，通过在自己的环境中管理生成式 AI 模型，您可以控制访问权限、简化法规遵从性、增强数据安全性，并在整个 AI 解决方案中保持更高的透明度。

最后，IBM Granite 系列模型提供了保障政策，若客户因所提供的开源软件或 AI 模型侵犯第三方知识产权而遭索赔，IBM 将承担相关索赔责任。在复杂多变的 AI 技术环境中，选择这类模型和供应商有助于进一步保护您的企业组织。

3　定制 AI 解决方案

开源 SLM 可助力快速高效地开发出贴合特定业务需求的 AI 解决方案。这些模型专为特定用例设计和构建，让您能够精准应对特定领域的挑战，同时规避通用 LLM 的复杂性和高资源需求。

通过利用企业数据对 SLM 进行调优，您可以将企业组织的专业知识和领域专长直接融入模型参数之中。这种方式能够提高 SLM 响应的相关性，降低重新训练的频率和成本，并缩短关键 AI 应用及服务的开发周期。

与 LLM 相比，SLM 规模较小且对数据要求较低，因此更易于定制，使您能够开发出针对特定任务或领域优化且精准高效的模型。在资源有限的环境和边缘部署中，SLM 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。

IBM Granite 模型之类的 SLM 还能简化从实验环境到生产环境的过渡流程。SLM 与各种硬件和软件基础架构实现简便集成，使您能够根据企业 IT 环境量身定制生成式 AI 解决方案。这种适应性有助于降低运维复杂性，同时保持对部署和性能的掌控。

4　降低 AI 模型成本

对于许多企业组织而言，降低 AI 的计算需求是有效管控开支的关键。与 LLM 相比，开源 SLM 在提供先进生成式 AI 解决方案所需性能的同时，还能降低训练和推理成本，并减少所需的计算能力。

SLM 的规模通常比主流 LLM 小数千倍，因此所需的计算资源、数据和能源要少得多。这种高效性有助于缩短训练时间、简化微调流程，并为 AI 开发提供更具可持续性的方法。

此外，开源 SLM 能够在多个项目和企业组织中高效扩展，而无需进行成本高昂的硬件升级。通过在现有 IT 基础架构中部署这些模型，您可以创建定制化的 AI 解决方案，而不会影响性能或超出预算限制。

成本节省不仅体现在基础架构方面。开源 SLM 还免除了与专有模型相关的许可费用，让您能够以经济实惠的方式获取先进的生成式 AI 功能，且不受供应商施加的各种限制约束。

红帽与 IBM 强强联手，以开源 SLM 助力创新

Granite 系列开源生成式 AI 模型由 IBM 开发，并已纳入红帽® 企业 Linux® AI，可满足企业级 AI 应用的特定需求。

进一步了解开源生成式 AI 模型

阅读利用开源模型，释放 AI 创新潜力电子书，了解有关适用于特定任务的 SLM 和开源生成式 AI 解决方案的更多信息。

关于红帽

红帽致力于帮助客户跨环境实现标准化，助力开发云原生应用，并利用红帽一流的支持、培训和咨询服务，实现复杂环境的集成、自动化、安全防护和管理。

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

cn.redhat.com
0225_KVM