亚马逊云科技Amazon Bedrock全新升级

Amazon Bedrock新增自动化推理检查、多智能体协作和模型蒸馏三项新功能，基于坚实的企业级功能基础构建，助力客户更快地从概念验证过渡到生产级的生成式人工智能

北京2024年12月12日 /美通社/ -- 亚马逊云科技在2024 re:Invent全球大会上，宣布推出Amazon Bedrock的三项新功能。Amazon Bedrock是一项完全托管的服务，借助高性能基础模型，构建和扩展生成式人AI应用程序。新发布的功能可帮助客户防止因模型幻觉而导致的事实性错误，编排多个AI智能体处理复杂任务，以及创建更小的、针对特定任务的模型，这些模型在成本与延迟方面远低于大型模型的同时，仍能达成相近效果。

自动化推理检查功能是强有力的生成式AI保护措施，有助于防止因模型幻觉而导致的事实性错误，从而开辟了需要更高精度的新的生成式AI用例。 多智能体协作功能可帮助客户轻松构建和编排多个AI智能体，以共同解决问题，从而扩展客户应用生成式AI解决复杂用例的方式。 模型蒸馏功能使客户能够将特定知识从功能强大的大模型转移到更小、更高效的模型，运行速度最快可提高500%，成本降低75%。 如今，数以万计的客户在使用 Amazon Bedrock，穆迪公司、普华永道和Robin AI等公司都在使用这些新功能，以经济高效的方式扩大推理规模，实现前所未有的生成式AI创新。

亚马逊云科技人工智能和数据副总裁Swami Sivasubramanian博士表示："Amazon Bedrock拥有广泛的模型选择、领先的功能，使开发人员能够更轻松地将生成式AI集成到其应用中，并且注重安全和隐私。对于希望将生成式AI作为其应用和业务核心的客户来说，Amazon Bedrock已成为一个不可或缺的工具。因此，Amazon Bedrock的客户群仅在去年就增长4.7倍之多。随着生成式AI逐渐改变越来越多企业业务和客户体验，推理将成为每个应用的核心部分。随着此次新功能的推出，我们正在为客户进行创新，以解决整个行业在将生成式AI应用推向生产时面临的主要挑战，比如降低幻觉和成本。"

自动化推理检查功能防止因模型幻觉而导致的事实性错误

虽然模型在不断进步，但即使是能力最强的模型也会产生幻觉，提供不正确或误导性的响应。幻觉仍然是整个行业面临的一个根本挑战，这限制了企业对生成式AI的信任。特别是在医疗保健、金融服务和政府机构等受监管的行业中，准确性至关重要，这些组织需要进行审核，以确保模型做出适当的响应。自动化推理检查功能是强大的生成式AI保护措施，可通过逻辑准确且可验证的推理来帮助防止因模型幻觉而导致的事实性错误。通过提高客户对模型响应的信任，自动化推理检查功能为生成式AI开辟了对准确性要求极高的新应用场景。

自动化推理是AI的一个分支，它运用数学来验证事情的正确性。在处理用户需要精确答案的问题时，自动化推理表现出色，尤其是在那些主题广泛且复杂、并有一套明确定义的规则或知识体系的领域。亚马逊云科技拥有一支由世界一流的自动化推理专家组成的团队，他们过去十年使用这项技术在整个亚马逊云科技改善用户体验，例如准确部署证明权限和访问控制以增强安全性，或者在部署之前，对Amazon Simple Storage Service（Amazon S3）中的数百万个场景进行检查，以确保可用性和持久性得到保障。

Amazon Bedrock Guardrails使客户能够轻松地将安全和负责任的AI检查应用到生成式AI应用程序中，从而指导模型仅讨论相关主题。通过Amazon Bedrock Guardrails，自动化推理检查功能可以让Amazon Bedrock验证事实响应的准确性，生成可审计的输出，并向客户准确展示模型得出结果的原因。这提高了透明度，并确保模型响应符合客户的规则和政策。例如，健康保险提供商采用了生成式AI驱动的客户服务应用程序，它需要能够正确响应客户有关保单的问题，自动化推理检查功能可实现这一点。为了使用该功能，提供商无需自动化推理方面的专业知识，只需上传其政策信息，Amazon Bedrock 会自动制定必要的规则，并指导客户反复测试，以确保模型调整为正确的响应。然后，保险提供商应用自动化推理检查功能，当模型生成响应时，Amazon Bedrock 会对其进行验证。如果响应不正确，例如弄错了免赔额或标记了不在承保范围内的程序，Amazon Bedrock 会使用自动化推理检查功能中的信息来建议正确的响应。

全球专业服务公司普华永道正在使用自动化推理检查功能来创建高度准确、可信且有用的AI助手和智能体，以推动其客户的业务处于领先地位。普华永道将该功能纳入到为金融服务、医疗保健和生命科学领域客户提供的特定行业解决方案，包括验证AI生成的合规内容是否符合美国食品药品管理局（FDA）和其他监管标准的应用程序。在公司内部，普华永道采用自动化推理检查功能来确保生成式AI助手和智能体生成的响应准确且符合内部政策。

轻松构建和协调多个智能体以执行复杂的工作流程

随着企业将生成式AI作为其应用程序的核心部分，这项技术的应用不再仅限于总结内容和增强聊天体验。企业还希望自己的应用程序能够执行实际操作。AI驱动的智能体可以通过利用模型的推理功能，将任务（例如帮助退货或分析客户留存数据）分解为模型可以执行的一系列步骤，从而帮助客户的应用程序完成这些操作。Amazon Bedrock智能体功能使客户能够轻松构建智能体，使其能够跨公司系统和数据源工作。单个智能体可能很有用，但更复杂的任务，如对数百或数千个不同变量进行财务分析，可能需要大量各具专长的智能体。然而，要创建一个能够协调多个智能体、在智能体之间共享上下文并动态分配不同任务给相应智能体的系统，需要专业工具和生成式AI专业知识，这是很多企业难以企及的。因此，亚马逊云科技扩展Amazon Bedrock智能体功能以支持多智能体协作，使客户能够轻松地构建和协调专业智能体来执行复杂的工作流程。

凭借Amazon Bedrock多智能体协作功能，客户可以通过为项目的特定步骤创建和分配专用智能体，从而获得更准确的结果，并通过编排多个并行工作的智能体来加速任务。例如，金融机构可以使用Amazon Bedrock智能体对一家公司进行投资前的尽职调查。首先，客户可以使用 Amazon Bedrock智能体创建一系列专注于特定任务的专用智能体，例如分析全球经济因素、评估行业趋势和审查公司的历史财务状况。在创建完所有专用智能体后，再创建一个主管智能体来管理整个项目。主管智能体负责协调工作，包括将任务分解并路由到相应的智能体，为特定智能体提供完成工作所需的信息，并确定哪些操作可以并行处理，以及哪些操作需要等待其他任务的详细信息完成后才能继续。一旦所有专业智能体都完成了自己的任务，主管智能体会将信息汇总，综合结果，并制定整体风险概况。

穆迪公司是信用评级和金融洞察领域的全球领导者，已选择Amazon Bedrock多智能体协作功能来增强其风险分析工作流程。穆迪公司正在利用Amazon Bedrock创建智能体，为每个智能体分配特定的任务，并允许其访问量身定制的数据集，以履行其职责。例如，一个智能体可能会分析宏观经济趋势，另一个智能体可能会使用专有财务数据评估公司特定风险，第三个智能体则考虑竞争和战略定位。这些智能体无缝协作，将输出结果综合成精确、可操作的洞察。这种创新方法使穆迪公司能够提供更快、更准确的风险评估，巩固其作为金融决策领域值得信赖的权威机构的声誉。

使用模型蒸馏功能创建更小、更快、更具成本效益的模型

如今，客户正在尝试各种型号的模型，以找到最适合其业务独特需求的模型。然而，即使在所有可用模型中，也很难找到一个能够提供特定知识、成本和延迟的最佳组合。较大的模型知识更丰富，但响应时间更长、成本更高，较小的模型运行速度更快、更便宜，但功能不够强大。模型蒸馏是一种将知识从大型模型转移到小型模型，同时保留小型模型性价比的技术。然而，这项工作需要机器学习（ML）专业知识来处理训练数据、手动微调模型，以及在不损害客户最初选择该小型模型的性能特征的前提下调整模型权重。借助Amazon Bedrock模型蒸馏功能，任何客户现在都可以蒸馏出自己的模型，与原始模型相比，被蒸馏模型的速度可以提高500%，运行成本降低75%，在检索增强生成（RAG）等用例中，准确性损失低于2%。现在，客户无需具备专业的机器学习知识，可根据自身用例进行优化，实现功能、准确性、延迟和成本的最佳组合。

借助Amazon Bedrock模型蒸馏功能，客户只需为给定的用例选择最佳模型，该系列模型中的一个较小的模型便能满足客户对成本和延迟的需求。在客户提供样本提示后，Amazon Bedrock将完成生成响应和微调较小模型的所有工作，如果需要，它甚至可以创建更多样本数据以完成蒸馏过程。这为客户提供了一个具有大模型的相关知识和准确性，同时又具有较小模型的速度和成本的模型，使其成为生产场景下（如实时聊天互动）的理想选择。模型蒸馏适用于来自Anthropic、Meta的模型，以及新发布的Amazon Nova模型。

Robin AI提供一款AI驱动的助手，该服务能使复杂的法律流程更快、更经济且更易于访问。该公司正在使用模型蒸馏来帮助实现针对数百万合同条款的高质量法律问答。模型蒸馏帮助Robin AI以极低的成本获得所需的准确性，而更快的响应则为客户与AI助手之间提供了更流畅的互动。

自动化推理检查、多智能体协作和模型蒸馏功能均已预览可用。

原文地址：http://www.china-nengyuan.com/news/218762.html