

亚马逊云科技独家首推Palmyra X5

Palmyra X5 是专为高效驱动多步骤 agents 而开发的模型，现仅可通过 Writer 和 Amazon Bedrock 以完全托管的方式提供。

北京 2025年4月29日 /美通社/ -- 亚马逊云科技宣布，在Amazon Bedrock上正式推出Palmyra X5——这是一款全新的具备一百万token上下文窗口的自适应推理模型。该模型由企业级生成式AI领军企业Writer发布，是首批在Amazon Bedrock提供如此大规模上下文窗口的模型之一。该模型在速度和成本效率方面进行了优化，使客户能够构建先进的多步骤AI agents，并能精确处理海量企业数据，从根本上改变了推理方式。亚马逊云科技现已成为首家且目前唯一提供Writer完全托管、无服务器模型的云提供商，包括最新Palmyra X5及Palmyra X4，其他更多模型也将很快上线。

随着生成式AI技术的加速发展，客户需要广泛的模型选择以精准匹配业务需求。Writer模型在Amazon Bedrock的推出，进一步丰富了Amazon Bedrock广泛的完全托管模型的选择，这些模型均来自领先的AI企业，帮助客户能更轻松、安全地构建和扩展生成式AI应用，推动业务转型与创新。

Palmyra X5是首批在Amazon Bedrock提供一百万token上下文窗口的模型之一，为亚马逊云科技的客户提供更多选择（上下文窗口指一个模型每次输入/请求能处理和"记忆"的信息量。它以模型处理的最小文本单位token为计量单位，可视为模型的"短期记忆"）。凭借这种规模的上下文窗口，Palmyra可精准处理1500页内容（相当于6本书）。该模型还是业界首批企业级自适应推理模型之一，并兼具先进的大语言模型能力以及扩展的记忆、处理能力。企业现可在预览范围内处理广泛任务，包括财务报告、法律合同分析、医疗记录整合、客户反馈挖掘等。

除推理能力外，Palmyra X5还具有多项强大功能，能够支持与系统交互的agents、支持高级代码生成与部署，并支持超过30种语言。

Palmyra X5 介绍：

若将Palmyra X5模型拟人化，它就像具有超能力——可在22秒内读完一百万个单词的内容，并即时生成可执行的见解。它不仅能完整记忆200页战略文档的全部内容，更能理解其与昨天的客户会议、上个季度财务数据的内在关联。面对复杂的问题时，它能系统化地逐步推进解决方案，全程清晰阐述思考路径——无论是帮助分析海量客户反馈以提炼共性，还是排查技术故障。

Writer 首席技术官兼联合创始人 Waseem AlShikh 表示："我们选择亚马逊云科技作为首家提供Writer完全托管模型的主流云服务提供商，源于其无与伦比的安全性，以及双方在变革企业AI应用方式、驱动创新增长愿景上的高度契合。Palmyra X5是Writer迄今最先进的模型，它能高速处理海量企业数据，这对于扩展多代理系统（multi-agent systems）至关重要。通过Amazon Bedrock，我们将这些强大功能带给全球更多企业，帮助客户在安全、可扩展的环境中部署。"

亚马逊云科技 Amazon Bedrock 总监 Atul Deo 表示："基于与Writer的深度战略合作，我们很高兴通过Amazon Bedrock提供Writer的Palmyra系列模型，赋能企业开启智能Agent创新的新时代。Palmyra X5在长上下文窗口中提供卓越性能，兼具企业级可靠性与速度。无缝接入Writer的Palmyra X5将让开发者和企业利用亚马逊云科技的安全性、扩展性和性能，构建和扩展AI agents，实现海量企业数据推理范式的革新。"

数据解析：

- Palmyra X5是最高效的大规模上下文大语言模型之一，对速度与成本均进行了优化。它可在约22秒内处理完整百万token提示词，单次函数调用响应仅需约0.3秒。
- 在最新Longbench v2评测中，Palmyra X5以53%的平均得分展现其同类领先的性价比。企业在大幅降低每百万token成本的同时，可获得接近顶级的准确性，在预算可控的情况下，可执行大量agents与长上下文处理任务。
- 支持30多种语言，为全球企业提供真正的多语言处理能力。
- 定价为每百万输入token 0.60美元，每百万输出token 6美元，是目前最具成本效益的大规模上下文大语言模型之一。
- 在BigCodeBench (完整版, 指令版)评估中，Palmyra X5以48.7分位列顶级模型，展现了其解决实用且具有挑战性的复杂编程任务的能力。

尽管生成式AI正在改变我们创建、分析和交互信息的方式，但Agentic AI将从根本上重塑工作本质。AI的这一新前沿超越了内容生成与洞察提炼，转向能够自主规划、执行和调整复杂动作序列的AI agents。通过Amazon Bedrock提供的Palmyra X5，亚马逊科技的客户可使用Writer的模型安全私密地构建和扩展AI agents，无需管理底层基础设施。此外，对各行业的企业来说，Palmyra X5最令人振奋的是能够构建和部署更复杂的AI agents，这些agents可以处理海量数据，并与其他agents、大语言模型及外部系统工具进行交互。

Writer提供精确且完全自主的模型，杜绝后期训练量化与知识蒸馏，确保今天验证的行为模式与明天的行为模式保持一致。Palmyra X5在此基础上强化技术承诺，严格保持后向兼容性以避免团队重复调优流程的痛苦，同时发布客户可参与制定的公开企业技术路线图，并通过优化推理延迟，使即便在百万token量级下，大语言模型交互和检索增强生成（RAG）也能实现近乎即时的响应。

Writer宣布，得益于创新型Transformer设计（一种支持输入数据并行处理而非顺序处理的架构）与混合注意力机制（允许多种方式同时关注信息，既保证效率也确保效果），其未来发布的所有大语言模型将以一百万token作为最小上下文窗口规模。这意味着企业可基于持续扩展的AI能力制定长期战略，而不受限于上下文窗口的规模约束。

原文地址：<http://www.china-nengyuan.com/news/225400.html>