AI 什么时候才能真正改变软件？我们对 Meta 的 Llama 4 发布感到非常兴奋，这是一个代表着开源 AI 模型的重要进步。我们非常高兴能成为 Meta 的合作伙伴，并从第一天起就支持 Llama 4。今天，我们很高兴地宣布，Llama 4 模型现已在 Together AI 平台上线。我们对 Llama 4 的性能感到非常兴奋，它在人类评估中表现出色，在多语言能力方面处于领先地位，并且在图像理解方面达到了业界领先水平。

Llama 4 Maverick 是一个拥有 17B 活跃参数和 128 个专家的 400B 参数模型，在 LMArena 上的聊天能力排名第二。Llama 4 Scout 是一个拥有 17B 活跃参数和 16 个专家的 109B 参数模型，具有业界领先的 10M 上下文窗口，并且在广泛的基准测试中表现优于 Gemma 3、Gemini 2.0 Flash-Lite 和 Mistral 3.1。

Llama 4 Maverick 和 Llama 4 Scout 现已在 Together AI 平台上提供无服务器 API，并且可以通过 Together 的专用端点进行企业规模的部署。Llama 4 Maverick 的价格为每百万输入/输出 token 0.27 美元/0.85 美元，Llama 4 Scout 的价格为每百万输入/输出 token 0.18 美元/0.59 美元。

Together AI 的推理引擎经过优化，即使在处理多模态和长上下文工作负载时，也能为 Llama 4 提供业界领先的性能。我们的 FP8 实现在保持高质量的同时，实现了高吞吐量和低延迟。

我们很高兴能够支持 Meta 最新的开源模型系列，并期待看到开发者和企业将 Llama 4 用于构建什么。

立即开始使用 Together AI 上的 Llama 4 API，或联系我们的销售团队，讨论企业部署选项。

Together AI 的 Llama 4 Maverick 和 Llama 4 Scout 现已上线，并提供业界领先的性能。

Llama 4 Maverick 是一个拥有 17B 活跃参数和 128 个专家的 400B 参数模型，在 LMArena 上的聊天能力排名第二。

Llama 4 Scout 是一个拥有 17B 活跃参数和 16 个专家的 109B 参数模型，具有业界领先的 10M 上下文窗口，并且在广泛的基准测试中表现优于 Gemma 3、Gemini 2.0 Flash-Lite 和 Mistral 3.1。

Llama 4 Maverick 和 Llama 4 Scout 现已在 Together AI 平台上提供无服务器 API，并且可以通过 Together 的专用端点进行企业规模的部署。