

硅基流动_SiliconCloud 生成式人工智能云服务平台-使用指南

硅基流动_SiliconCloud 生成式人工智能云服务平台提供的 DeepSeek 模型推理加速及部署的专属实例资源，依托于硅基流动高性能推理引擎平台，该引擎使用集中式调度器来协调分布式 GPU 工作器（Worker）的执行，多 GPU 节点、多种并行方式混合执行，并集成了 Contiguous Batching、KV Cache 和 Paged Attention、算子融合等多种优化技术，在设计过程中充分考虑了高性能、高可用性、易扩展性和安全性，确保在复杂多变的业务环境中，始终为用户提供高效、稳定、可靠的推理能力。

使用步骤

1. 使用您的华为云账号登录华为云商店
2. 提交购买申请
 - 点击页面右上方「立即购买」或「留言咨询」按钮
 - 确认订单信息，并完成支付
3. 产品顾问对接
 - 我们将在 5 个工作日内与您取得联系
 - 了解您的具体业务场景和技术需求
 - 提供针对性的部署方案与资源配置建议
4. 签署服务协议并准备部署
 - 明确所需 GPU 规格、服务周期、价格等
 - 签署合作协议并
 - 安排资源准备与部署计划

5. 启动部署服务

- 由硅基流动技术团队完成模型部署、接口配置与调优
- 部署周期视所选资源与模型而定

如有任何问题或需要进一步的帮助，欢迎邮件咨询：

support-huaweicloud@siliconflow.cn