可按需选择对应规格部署DeepSeek

天翼云 State Cloud
公有云解决方案中心

模型	推荐规格							
	通用云主机	T4	V100	V100S	A10	A100	H800	昇腾910B
DeepSeek-R1-1.5B	c7.2xlarge.2 (8C16G)	pi2.2xlarge.4 (显存：1*16G)	p2v.2xlarge.4 (显存：1*32G)	p2vs.2xlarge.4 (显存：1*32G)	pi7.4xlarge.4 (显存：1*24G)	p8a.6xlarge.4 (显存：1*40G)	1台 physical.h8ns.6xlarge28 (8*80G)	**1台** physical.acas910b.2xlarge11 (显存：4*8*64G)
DeepSeek-R1-7B	c7.4xlarge.2 (16C32G)	pi2.4xlarge.4 (显存：2*16G)	p2v.2xlarge.4 (显存：1*32G)	p2vs.2xlarge.4 (显存：1*32G)	pi7.4xlarge.4 (显存：1*24G)	p8a.6xlarge.4 (显存：1*40G)		
DeepSeek-R1-8B	-	pi2.4xlarge.4 (显存：2*16G)	p2v.2xlarge.4 (显存：1*32G)	p2vs.2xlarge.4 (显存：1*32G)	pi7.4xlarge.4 (显存：1*24G)	p8a.6xlarge.4 (显存：1*40G)		
DeepSeek-R1-14B	-	pi2.8xlarge.4 (显存：4*16G)	p2v.4xlarge.4 (显存：2*32G)	p2vs.4xlarge.4 (显存：2*32G)	pi7.8xlarge.4 (显存：2*24G)	p8a.12xlarge.4 (显存：2*40G)		
DeepSeek-R1-32B	-	-	p2v.8xlarge.4 (显存：4*32G)	p2vs.8xlarge.4 (显存：4*32G)	pi7.16xlarge.4 (显存：4*24G)	p8a.24xlarge.4 (显存：4*40G)		
DeepSeek-R1-70B	-	-	-	-	-	p8a.24xlarge.4 (显存：4*40G)		
DeepSeek-R1、 DeepSeek-V3满血版 (671B)	-	-	-	-	-	-		**4台** physical.acas910b.2xlarge11 (显存：4*8*64G)

备注：

1. 预装镜像:DeepSeek-Ubuntu22.04(预装DeepSeek-R1:7B): 推荐配置：内存≥8G、显存≥16G
2. 折扣价：普通云主机是按照25折进行计算，GPU云主机按照25折进行计算，可根据实际需求进行申请折扣；GPU裸金属为对省出货价，实际价格可走OA进行申请对应折扣
3. 此报价云主机只包含CPU、内存、GPU、系统盘，不包含EIP价格，网络需根据客户需求进行添加；GPU裸金属也不包含EIP价格

不同模型规模适应不同场景

模型规模	部署成本	准确性	适用场景	典型应用
1.5B-8B	较低	☆ ☆	学生、个人开发者、轻量级应用、小型团队等低成本快速试错需求	聊天机器人、本地文档分析、个人知识库、简单数据处理、文档辅助等
14B	中等	☆ ☆ ☆ 逻辑能力提升明显	小型企业、内容创作者	可处理上下文的高级智能客服、多轮复杂对话、长文总结、简单数据分析、写作助手等
32B	较高	☆ ☆ ☆ ☆ 专业领域明显增强	企业级服务/垂直领域	智能客服、代码生成、BI助手、企业知识库
70B	高	☆ ☆ ☆ ☆ ☆ 能力均衡，接近商用	科研机构/超复杂任务	药物研发、金融预测、AIGC生成
671B	极高	最强，推荐使用裸金属	国家级算力平台/前沿探索	气候模拟、通用人工智能研究

对应资源规格官网标准价格

资源类型	规格名称	vCPU(核)	内存(GB)	GPU	显存	官网原价(元/月)
普通云主机	c7.2xlarge.2	8	16	-	-	446
	c7.4xlarge.2	16	32	-	-	1060
GPU云主机	pi2.2xlarge.4	8	32	1×T4	1×16GB	3515
	pi2.4xlarge.4	16	64	2×T4	2×16GB	7030
	pi2.8xlarge.4	32	128	4×T4	4×16GB	14060
	p2v.2xlarge.4	8	32	1x V100	1x32GB	7377.8
	p2v.4xlarge.4	16	64	2*V100	2*32GB	14755.6
	p2v.8xlarge.4	32	128	4*V100	4*32GB	29511.2
	p2vs.2xlarge.4	8	32	1*V100s	1*32GB	7377.8
	p2vs.4xlarge.4	16	64	2*V100s	2*32GB	14755.6
	p2vs.8xlarge.4	32	128	4*V100s	4*32GB	29511.2
	pi7.4xlarge.4	16	64	1×A10	1×24GB	4447.43
	pi7.8xlarge.4	32	128	2×A10	2×24GB	8894.85
	pi7.16xlarge.4	64	256	4×A10	4×24GB	17789.69
	p8a.6xlarge.4	24	96	1×A100	1×40GB	10229.09
	p8a.12xlarge.4	48	192	2×A100	2×40GB	20458.17
	p8a.24xlarge.4	96	384	4×A100	4×40GB	40916.34
GPU裸金属	physical.h8ns.6xlarge28	2路48核48线程	2048	8xH800	8x80GB	237,500.00
	physical.acas910b.2xlarge11(风冷)	4路48核48线程	1536	8x昇腾910B	8x64GB	172,000.00
	physical.lcas910b.2xlarge11(液冷)	4路48核48线程	1536	8x昇腾910B	8x64GB	196,000.00