

LocalAI 产品系列规格

产品矩阵

系列版本	5070 系列	5090 系列	H20 系列
入门版 Essential	S-5070X4-64 4*RTX5070Ti/64GB		
进阶版 Pro	P-5070X8-128 8*RTX5070Ti/128GB	P-5090X4-128 4*RTX5090/128GB	P-H20X2-280 2*H20/282GB
企业版 Enterprise		E-5090X8-256 8*RTX5090/256GB	E-H20X4-560 4*H20/564GB
超算版 Cluster		U3-5090X24-768 3 节点 24*RTX5090/768GB	

5070 系列产品规格

型号	入门版 S-5070X4-64		进阶版 P-5070X8-128	
尺寸规格	4U		6U	
CPU	1 路 Intel Xeon Gold 6430 （支持 2 路 CPU）		2 路 Intel Xeon Gold 6430	
内存	DDR5-5600 768GB， 64GB × 12		DDR5-5600 768GB， 64GB × 12	
GPU	NVIDIA RTX 5070Ti 16GB × 4		NVIDIA RTX 5070Ti 16GB × 8	
PCIe 扩展	PCIe 5.0		PCIe 5.0	
系统硬盘	500GB SSD × 2（RAID 支持）		500GB SSD × 2（RAID 支持）	
数据硬盘	4TB SATA 3.0 × 1		4TB SATA 3.0 × 1	
网络	10GbE		10GbE	
电源	2000W		3000W	
操作系统	Ubuntu 22.04		Ubuntu 22.04	
部署模型	DeepSeek V3-0324 685B DeepSeek R1-0528 685B Qwen3-235B-A22B			
Throughput DeepSeek R1-0528 **I-1024/O-1024	1 并发	6.28 tokens/s	1 并发	9.80 tokens/s
	10 并发	9.73 tokens/s	10 并发	15.46 tokens/s

5090 系列产品规格

型号	进阶版 P-5090X4-128		企业版 E-5090X8-256	
尺寸规格	4U		6U	
CPU	1 路 Intel Xeon Gold 6530 （支持 2 路 CPU）		2 路 Intel Xeon Gold 6530	
内存	DDR5-5600 768GB， 64GB × 12		DDR5-5600 768GB， 64GB × 12	
GPU	NVIDIA RTX 5090 32GB × 4		NVIDIA RTX 5090 32GB × 8	
PCIe 扩展	PCIe 5.0		PCIe 5.0	
系统硬盘	500GB SSD × 2（RAID 支持）		500GB SSD × 2（RAID 支持）	
数据硬盘	4TB SATA 3.0 × 1		4TB SATA 3.0 × 1	
网络	10GbE		10GbE	
电源	3000W		5500W	
操作系统	Ubuntu 22.04		Ubuntu 22.04	
部署模型	DeepSeek V3-0324 685B DeepSeek R1-0528 685B Qwen3-235B-A22B			
Throughput DeepSeek R1-0528 **I-1024/O-1024	1 并发	7.64 tokens/s	1 并发	10.50 tokens/s
	50 并发	35 tokens/s	50 并发	60 tokens/s