

硬件

GPU 需求表

模型	参数量	显存需求	性能需求	适用场景
DeepSeek 7B	7B	16GB+	单卡可运行	轻量任务（聊天、代码补全）
DeepSeek 32B	32B	48GB+	2~4 张 24GB+ GPU	适合问答、代码分析
DeepSeek 70B	70B	80GB+	4~8 张 80GB GPU	高级推理、复杂任务
DeepSeek 67B MoE	67B MoE	160GB+	8 张 80GB GPU	最高端，适合科研和高精度 NLP

显卡型号 & 价格对比

显卡型号	显存 (GB)	算力 (TFLOPS FP16)	功耗 (W)	适用模型	单卡价格 (¥)	多卡配置价格 (¥)
RTX 4090	24GB	82 TFLOPS	450W	7B	¥ 16,000	¥ 16,000 (单卡可跑 7B)
RTX 3090	24GB	36 TFLOPS	350W	7B / 32B (多卡)	¥ 8,000	¥ 32,000 (4 卡可跑 32B)
RTX 6000 Ada	48GB	91 TFLOPS	300W	32B	¥ 55,000	¥ 110,000 (2 卡可跑 32B)
A6000	48GB	39 TFLOPS	300W	32B	¥ 30,000	¥ 120,000 (4 卡可跑 32B)
A100 40GB	40GB	78 TFLOPS	400W	32B	¥ 45,000	¥ 90,000 (2 卡可跑 32B)

显卡型号	显存 (GB)	算力 (TFLOPS FP16)	功耗 (W)	适用模型	单卡价格 (¥)	多卡配置价格 (¥)
A100 80GB	80GB	78 TFLOPS	400W	70B	¥ 90,000	¥ 360,000 (4 卡可跑 70B)
H100 80GB	80GB	197 TFLOPS	700W	67B MoE	¥ 250,000	¥ 2,000,000 (8 卡可跑 67B MoE)

预算 & 适配推荐

预算范围 (¥)	适用模型	推荐 GPU 方案
< 3 万	7B	1 × RTX 4090
5~20 万	32B	2~4 × RTX 6000 / A6000
50~100 万	70B	4~8 × A100 80GB
200 万以上	67B MoE	8 × H100 80GB

选择建议

1. 预算 3 万以内：选 RTX 4090（能跑 7B，适合个人/小团队）。
2. 预算 10~20 万：选 2~4 张 RTX 6000 / A6000（能跑 32B）。
3. 预算 50~100 万：选 4~8 张 A100 80GB（能跑 70B，适合企业）。
4. 预算 200 万以上：选 8 张 H100 80GB（能跑 67B MoE，适合大厂和科研）。

部署

```
docker image load -i ./recognize_general_receipt_21398_cuda11.gpu.tar
docker run --ulimit core=0 --gpus all --net=host --restart=always -v /opt/hh/ocr-
log/general_receipt:/recognize_document/log -d
recognize_general_receipt_21398_cuda11:gpu -p 8015 -I -G 0,0,0,0,0 -L 172.24.72.225
```

```
[root@fe-ocr-gpu4 ~]# docker ps
```

CONTAINER ID	IMAGE	COMMAND
CREATED	STATUS	PORTS
		NAMES
9e4e07f9b365	recognize_general_receipt_21398_cuda11:gpu	
"./recognize_documen..."	3 days ago	Up 3 days
hungry_bhaskara		

```
(base) [root@fe-ocr-gpu5 ~]# ollama list
```

NAME	ID	SIZE	MODIFIED
linux6200/bge-reranker-v2-m3:latest	abf5c6d8bc56	1.2 GB	8 days ago
deepseek-r1:70b	0c1615a8ca32	42 GB	2 weeks ago
bge-m3:latest	790764642607	1.2 GB	3 weeks ago
deepseek-r1:32b	38056bbcbb2d	19 GB	3 weeks ago
qwen2.5:32b	9f13ba1299af	19 GB	4 weeks ago

训练