AI 大模型部署方案成本对比分析

自建机房 vs GPU云租赁 vs API调用 | 2026年6月

一、不同模型对 GPU 的需求

部署成本首先取决于模型规模，下表列出了当前主流模型的 GPU 配置需求和硬件采购价格参考。

模型规模	代表模型	最低显存	推荐GPU配置	硬件采购价
7B~14B	Qwen3-8B、DeepSeek-14B	16~32 GB	1×A100-40G / 1×L40S	¥5~15万
32B~70B	Qwen3-32B、Llama3-70B	64~128 GB	2~4×A100-80G / 2~4×H100	¥30~100万
70B~235B	Qwen3-235B (MoE)	160 GB+	4~8×A100-80G / 4~8×H100	¥50~200万
671B (满血)	DeepSeek-V3/R1 完整版	320 GB+	8×A100-80G / 8×H100	¥80~200万

二、方案一：API 调用

API调用是最轻量的方案，零固定成本，按量付费。2026年DeepSeek掀起的价格战使国内API均价较2023年累计下降超过90%。

主流API价格（2026年6月，元/百万Tokens）

厂商	模型	输入价格	输出价格	定位
DeepSeek	V4-Flash	0.20	0.42	极致性价比
DeepSeek	V4-Pro	3	6	主力旗舰
DeepSeek	V3	2	8	均衡之选
阿里	Qwen-Plus	0.8	2	中文高性价比
阿里	Qwen3-Max	6	24	中文旗舰
字节	Doubao-pro-32k	0.8	2	轻量场景
智谱	GLM-4-Air	1	1	学术/知识库
Kimi	K2.5	~2.8	~4	长文本
OpenAI	GPT-4o	18	72	国际标杆
Anthropic	Claude Sonnet 4.6	21	105	Agent/编程最强

月度API费用估算（日活1万用户，人均800输入+400输出tokens/天）

模型	月Token量	月费
DeepSeek-V4-Flash	2.4亿输入 + 1.2亿输出	¥98
Qwen-Plus	同上	¥432
DeepSeek-V3	同上	¥1,440
DeepSeek-V4-Pro	同上	¥1,440
Qwen3-Max	同上	¥4,320
GPT-4o	同上	¥12,960
Claude Sonnet 4.6	同上	¥17,640

三、方案二：GPU 云租赁

在公有云上租GPU实例，自己部署开源模型（如vLLM/TGI），省去机房建设，电费散热由云厂商承担，仍需1名运维人员负责模型部署与调优。

GPU云端租赁价格（2026年Q2）

GPU	显存	按需(元/小时)	预留1年(元/月)	年费(预留)
A100 80GB	80 GB	~¥14	~¥6,500	~¥7.8万
H100 80GB	80 GB	~¥25	~¥11,500	~¥13.8万
H200 141GB	141 GB	~¥32	~¥15,000	~¥18万
L40S 48GB	48 GB	~¥7	~¥3,200	~¥3.8万

典型配置月费

场景	GPU配置	吞吐量	月费(预留)
轻量推理 (7B)	1×A100-80G	~180 tok/s	¥6,500
中等推理 (32B)	2×A100-80G	~360 tok/s	¥13,000
高并发 (70B+)	4×H100-80G	~720 tok/s	¥46,000
满血671B推理	8×H100-80G	~1,440 tok/s	¥92,000

四、方案三：自建机房

全私有化部署，数据完全可控。成本由一次性建设投入+持续运维支出组成。

一次性建设投入

规模	GPU配置	硬件	基建	软件+实施	建设总计
轻量级	1~2×A100-40G	¥8~15万	¥5~10万	¥3~5万	¥16~30万
标准级	2~4×A100-80G	¥30~70万	¥15~25万	¥5~10万	¥50~105万
企业级	4~8×H100-80G	¥80~200万	¥25~50万	¥10~20万	¥115~270万
集群级	8+×H100/集群	¥200~500万+	¥50~100万+	¥20~50万+	¥270~650万+

年度运维支出

项目	轻量级	标准级	企业级
电费 (GPU+空调)	¥3~5万	¥8~15万	¥20~40万
运维人员 (1~2人)	¥20万	¥30万	¥40~60万
硬件维保/带宽	¥2~3万	¥5~8万	¥10~20万
年运维合计	¥25~28万	¥43~58万	¥70~120万

GPU购买 vs 云租赁盈亏平衡

GPU	购买价	云租赁年费	盈亏平衡点
A100 80GB	¥10万	¥7.8万/年	~15个月
H100 80GB	¥22万	¥13.8万/年	~19个月
H200 141GB	¥30万	¥18万/年	~20个月

结论：如果连续使用超过1.5~2年，购买GPU比云租赁更划算。

五、三年 TCO 总成本横评

场景A：轻量推理 (7B~14B模型，知识库问答/客服)

方案	第一年	第二年	第三年	三年总计
API (Qwen-Plus)	¥0.5万	¥0.5万	¥0.5万	¥1.6万

API (DeepSeek-V4-Flash)	¥0.1万	¥0.1万	¥0.1万	¥0.4万
GPU云租赁 (1×A100)	¥7.8万	¥7.8万	¥7.8万	¥23.4万
自建机房	¥41~58万	¥25万	¥25万	¥91~108万

轻量场景 API 方案碾压性胜出，三年费用仅为自建的 1/100。

场景B：中等负载 (32B~70B模型，日活3万，高并发推理)

方案	第一年	第二年	第三年	三年总计
API (DeepSeek-V3)	¥17万	¥17万	¥17万	¥52万
API (DeepSeek-V4-Pro)	¥17万	¥17万	¥17万	¥52万
GPU云租赁 (2×A100)	¥15.6万	¥15.6万	¥15.6万	¥47万
自建机房	¥95~148万	¥43万	¥43万	¥181~234万

API 与云租赁接近打平。若吞吐量稳定，云租赁略优；若波动大，API弹性更好。

场景C：重度负载 (70B+模型，日活10万+，多业务并发)

方案	第一年	第二年	第三年	三年总计
API (DeepSeek-V4-Pro)	¥173万	¥173万	¥173万	¥518万
GPU云租赁 (8×H100)	¥110万	¥110万	¥110万	¥331万
自建机房	¥360~510万	¥70万	¥70万	¥500~650万

GPU云租赁三年最优；自建机房在3~4年左右开始显现成本优势。

六、决策矩阵

判断条件	推荐方案
日活<5000，轻量模型即可	API调用 (DeepSeek/Qwen-Plus)，月费仅数百元
日活5000~5万，对延迟敏感	GPU云租赁，自管模型获得更优吞吐
日活>5万且稳定，有数据合规硬需求	自建机房，3~4年回收
业务波动大、快速试错阶段	API调用，零固定成本，随时切换模型
金融/医疗/政务等高合规要求	自建或云租赁，数据绝不外传
需要定制微调专属模型	云租赁（灵活）→ 成熟后考虑自建

七、混合方案：最务实的路径

多数企业最优解并非三选一，而是分层混合：

1. 上线初期：全部走API，验证PMF，月费几乎可忽略。
2. 稳定增长期：高频场景切到云租赁（1~2张GPU），低频保留API。
3. 规模化阶段：核心业务自建小型GPU集群，长尾场景继续用API。

典型路径：先用DeepSeek API（¥100/月）跑通业务 → 切部分负载到租用A100

→ 业务成熟后自建2~4卡服务器，三年总成本可控制在50~100万以内。