1.5B 模型（约15亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 4GB（需4-bit量化） | 8GB（可运行FP16精度） |
| 显卡型号 | NVIDIA GTX 1650/1060 6GB | NVIDIA RTX 3060/4060 8GB |
| 内存（RAM） | 8GB DDR4 | 16GB DDR4 |
| CPU | Intel i3 / AMD Ryzen 3（4核） | Intel i5 / AMD Ryzen 5（6核） |
| 存储 | 256GB SSD（模型文件约3-5GB） | 512GB NVMe SSD |
| 适用场景 | 轻量文本生成、简单问答、嵌入式设备 | 本地调试、小型应用部署 |

7B 模型（约70亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 8GB（需4-bit量化） | 16GB（FP16精度） |
| 显卡型号 | RTX 3060/3070/4060 8GB | RTX 3080/4080/A2000 16GB |
| 内存（RAM） | 16GB DDR4 | 32GB DDR5 |
| CPU | Intel i5 / Ryzen 5（6核） | Intel i7 / Ryzen 7（8核） |
| 存储 | 512GB SSD（模型文件约10-15GB） | 1TB NVMe SSD |
| 适用场景 | 本地对话、代码补全、中等长度生成 | 多任务推理、长文档处理 |

8B 模型（约80亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 10GB（需4-bit量化） | 16-24GB（FP16精度） |
| 显卡型号 | RTX 3080 10GB/4060 Ti 16GB | RTX 3090/4090/A4000 16GB+ |
| 内存（RAM） | 24GB DDR4 | 48GB DDR5 |
| CPU | Intel i7 / Ryzen 7（8核） | Intel i9 / Ryzen 9（12核） |
| 存储 | 512GB SSD（模型文件约12-18GB） | 1TB NVMe SSD |
| 适用场景 | 复杂对话、多轮交互、中等规模数据分析 | 企业级工具开发、RAG应用 |

14B 模型（约140亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 16GB（需4-bit量化） | 24GB+（FP16精度） |
| 显卡型号 | RTX 3090 24GB/4090 24GB | NVIDIA A5000 24GB/A6000 48GB |
| 内存（RAM） | 32GB DDR4 | 64GB DDR5 |
| CPU | Intel i9 / Ryzen 9（12核） | Xeon/EPYC（16核以上） |
| 存储 | 1TB NVMe SSD（模型文件约25-30GB） | 2TB NVMe SSD |
| 适用场景 | 代码生成、复杂逻辑推理、企业级工具 | 高并发API服务、大规模数据处理 |

32B 模型（约320亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 24GB（需4-bit量化 + 多卡） | 80GB+（FP16精度 + 多卡） |
| 显卡型号 | 2x RTX 3090 24GB | 2x NVIDIA A100 80GB/H100 80GB |
| 内存（RAM） | 64GB DDR4 | 128GB DDR5 ECC |
| CPU | Xeon/EPYC（16核以上） | 双路CPU（32核以上） |
| 存储 | 2TB NVMe SSD（模型文件约60-80GB） | 4TB NVMe RAID |
| 适用场景 | 科研级推理、大规模知识库问答 | 分布式训练、超长文本生成 |

70B 模型（约700亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 48GB（需4-bit量化 + 4卡） | 320GB+（FP16精度 + 多卡集群） |
| 显卡型号 | 4x RTX 4090 24GB | 4x NVIDIA H100 80GB/A100 80GB |
| 内存（RAM） | 128GB DDR5 ECC | 256GB+ DDR5 ECC |
| CPU | 双路Xeon/EPYC（64核以上） | 四路CPU（128核以上） |
| 存储 | 4TB NVMe RAID（模型文件约140GB） | 8TB 企业级SSD阵列 |
| 适用场景 | 类GPT-3.5级别推理、超大规模数据处理 | 云端服务、AIaaS平台 |

671B 模型（约6710亿参数）

| 配置项 | 最低配置 | 推荐配置 |
| --- | --- | --- |
| 显存（GPU） | 无单机方案，需分布式集群 | 64x NVIDIA H100（6400GB显存） |
| 显卡组合 | 云服务（AWS P4/P5实例） | 超算集群（千卡级并行） |
| 内存（RAM） | 512GB+ ECC DDR5 | 2TB+ ECC DDR5 |
| CPU | 多路EPYC/Xeon（256核以上） | 超算级CPU（1024核以上） |
| 存储 | 16TB+ 高速存储阵列 | 分布式存储（100TB+） |
| 适用场景 | 国家级AI研究、超大规模预训练 | 全球级AI服务（如GPT-4级别） |

通用优化策略

1.5B~14B：优先使用 4-bit量化（QLoRA）降低显存占用。适合个人开发者或中小企业，中端硬件可流畅运行。

32B+：需结合 模型并行 + 8-bit量化。

70B+模型可通过 CPU卸载（如LLAMA.cpp）运行，但速度显著下降。需企业级硬件或云资源，成本高昂。

671B仅限国家级实验室或超大型企业，普通用户推荐API调用。