基本参数假设

1B 参数数量:（10亿个参数）

参数精度：

FP32：4字节/参数

FP16：2字节/参数

INT4（量化）：0.5字节/参数

优化器状态： 通常需要存储 2-3 倍的模型参数（比如动量、梯度累积等）。

LoRA 插入参数比例： 假设适配器权重占模型总参数的 0.1%。

QLoRA 分页机制： 假设所有基础模型参数都存储在 CPU，而显存仅存储 LoRA 模块和部分中间梯度。

(1) Full Finetuning（全量微调）

模型参数（FP16）：

1B×2bytes=2GB

优化器状态（FP16 动量+梯度等，假设 3 倍参数量）：

2GB×3=6GB

总内存需求：

2GB+6GB=8GB

(2) LoRA 微调模型参数（FP16，冻结，不更新）：

1B×2 bytes=2 GB

LoRA 适配器参数（假设 0.1% 参数量，FP16）：

1B×0.001×2 bytes=0.002 GB (即 2MB)

优化器状态（仅适配器，假设 3 倍）：

0.002GB×3=0.006GB(即 6MB)

总内存需求：

2GB+0.002GB+0.006GB=2.008GB

(3)QLoRA 微调

模型参数（INT4，存储在 CPU）：

1B ×0.5bytes=0.5GB

LoRA 适配器参数（0.1%，FP16）：

1B×0.001×2bytes=0.002GB

优化器状态（仅适配器，假设 3 倍）：

0.006GB

中间缓存（用于分页，假设 GPU 仅加载一部分参数）：

0.5GB×10%=0.05GB

总显存需求（GPU 部分）：

0.002GB+0.006GB+0.05GB=0.058GB

总 CPU 存储需求：

0.5+0.058=0.558GB