Ideas

https://arxiv.org/abs/2210.17323
在 GPTQ 这篇论文的基础上：

1. 做 Inference quantization
    - GPTQ 只进行了 weight quantization，只对存储的 weight 进行量化，运算过程中仍然是全精度，因而性能提升仅来自于更小的显存需求
    - Inference quantization 能进一步降低运算需求
2. 做 sparse quantization
    - 只对对结果影响比较小的 weights 做 quantization
    - 重要的 weights 不进行 quantization
    - 实现可能比较复杂，可能要魔改 kernel