量子化に関する実証的分析と知見

- アクティベーションの量子化は、モデルが大きくなると外れ値が発生して困難になる
- 量子化されたLLMの性能を向上するには、効率的なファインチューニングを行うのがよい
- 8ビットの量子化ではまず性能は低下しないが、4ビットや3ビットの量子化については、性能が下がらないような戦略が必要