知識蒸留

概要:

- 大規模モデル（教師）から小規模モデル（生徒）への知識転移技術
- 計算コストを抑えながら性能を維持する圧縮手法
- より効率的なモデルデプロイを実現

例: llama3.1 (8B) → llama3.2 (1B)

- より小さいモデルで同等性能
- 推論速度2倍向上
- メモリ使用量50%削減