



quantize\_model

PerLayerQuantizeModelTransformer

\_do\_quantization\_split

get\_layers\_quantize\_group\_map

PerLayerQuantizeLayoutTransform

PerLayerNBitQuantizeRegistry



quantize\_annotate\_model

PerLayerNBitQuantizeScheme



quantize\_apply

Model s falešnou kvantizací s bitovou  
přesností dle per-layer konfigurace

