Promp Length=2048, Decode Phase CPU: Op Distribution Over Model Run

- Transpose - 33.23 %
- Concat - 20.52 %
- MatMulNBits - 18.78 %
- Mul - 10.92 %
- MatMul - 7.61 %
- UNKNOWN--Start/Stop Overhead? - 4.07 %
- Slice - 1.13 %
- Gather - 0.72 %
- Add - 0.70 %
- Unsqueeze - 0.51 %
- Reshape - 0.34 %
- Shape - 0.28 %
- SimplifiedLayerNormalization - 0.26 %
- Div - 0.22 %
- Softmax - 0.16 %
- Neg - 0.15 %
- QuickGelu - 0.13 %
- Sqrt - 0.13 %
- Cast - 0.09 %
- Where - 0.01 %
- Expand - 0.01 %
- Sub - 0.01 %
- ConstantOfShape - 0.00 %
- Equal - 0.00 %
- Range - 0.00 %
- Less - 0.00 %
- Squeeze - 0.00 %